Enseignement du lexique assisté par ordinateur

Jurij D. Apresjan

Institut pour les problèmes de transmission d'information, Académie des Sciences de Russie, Moscou, Russie

Introduction

L'objectif final de ce projet est de créer un manuel multilingue sur PC destiné à enseigner le lexique des langues étrangères à travers le lexique de la langue maternelle et à apprendre le bon usage (correct et idiomatique) de ce dernier. Ce manuel s'adresse aux étudiants des écoles secondaires, des collèges, des universités, ainsi qu'aux personnes qui veulent individuellement apprendre une langue étrangère. Il leur permettra de se familiariser avec le lexique d'une langue étrangère ou de leur langue maternelle grâce aux techniques linguistiques originales basées sur les théories linguistiques modernes.

Dans le prototype que nous présentons aujourd'hui et qui fonctionne sur un IBM PC 486, seules deux langues sont accessibles : le russe et l'anglais. La taille des dictionnaires est également assez réduite : ils ne comportent qu'un millier de lexèmes dans chacune des langues. Comme on fournit systématiquement, pour chaque lexème, au moins un équivalent dans l'autre langue, nous avons pu garder un assez bon isomorphisme entre les deux dictionnaires. Actuellement, dans le cadre d'un projet INTAS, nous travaillons à la création d'un dictionnaire allemand. Dans les prochaines années, il est prévu d'y rajouter la version française et de porter la taille des dictionnaires à 3 000, voire 4 000 unités lexicales.

Sur le plan théorique notre manuel est essentiellement basé sur la théorie « Sens-Texte » de Mel'čuk (STM), dont il reprend les principaux concepts et notamment les fonctions lexicales (FL), le système de paraphrasage et le dictionnaire explicatif et combinatoire (DEC). Il s'appuie également sur le métalangage sémantique conçu par Apresjan pour décrire la sémantique des langues naturelles et sur sa théorie de la décomposition lexicale (définitions).

Le présent article se décompose en plusieurs sections : dans un premier temps, nous allons passer en revue quelques concepts généraux de la théorie STM (section

1), ensuite nous nous concentrerons sur la notion des fonctions lexicales de Mel'čuk (section 2), dans la 3e section nous montrerons comment les principes de la décomposition sémantique ont été mis en œuvre lors de la rédaction de notre manuel, enfin la 4e section décrit les « jeux linguistiques » sur PC issus des théories indiquées ci-dessus.

1. Concepts généraux de la théorie STM

Le modèle STM est un instrument logique destiné à établir pour une langue donnée des correspondances multiples entre les représentations sémantiques et les textes dans cette langue. Ce modèle s'oriente beaucoup plus vers l'expression que vers la compréhension, c'est-à-dire qu'il est plus adapté à la production des textes qu'à leur interprétation. Plus précisément, il permet de simuler trois aptitudes principales de la personne qui parle une langue quelconque, ces aptitudes caractérisant en fait la maîtrise de cette langue, à savoir :

a) l'aptitude à choisir les mots, les formes grammaticales et les constructions syntaxiques qui en principe pourraient être utilisés pour exprimer une pensée concrète. Par exemple, si l'on devait exprimer une pensée comme : « Le fait que la température de l'air ambiant est devenue soudain beaucoup plus basse a eu pour effet que les oiseaux nouveau-nés ont cessé de vivre ». Une personne parlant couramment l'anglais produirait sans difficulté un texte comme : *The nestlings died as a result of the cold wave* (Les couvées périrent à cause de l'arrivée d'un front froid).

b) l'aptitude à combiner correctement les mots, les formes grammaticales et les constructions syntaxiques retenus. On sait bien que les restrictions combinatoires ne s'expliquent pas toujours par des raisons sémantiques. Même les synonymes peuvent avoir un potentiel combinatoire différent. Ainsi, le mot anglais *word* (parole) dans une de ses acceptions est synonymique au mot *promise* (promesse) ; cf. *to give a promise <one's word> to do something* (donner la promesse <sa parole> de faire quelque chose) ; *to keep one's promise <one's word>* (tenir sa promesse <sa parole>) ; *to break one's promise <one's word>* (manquer à sa promesse <sa parole>). Cependant, seul le mot *promise* (promesse) peut s'employer avec les verbes *to make* et *to fulfil* (faire et remplir), cf. *to make a promise, to fulfil a promise* (faire une promesse, remplir la promesse), avec un adjectif numéral ordinal, cf. *three promises* (trois promesses) et avec un article défini ou indéfini, cf. *a promise, the promise* (une promesse, la promesse). Alors que *to make a word, to fulfil a word, three words, a word, the word* (faire une parole, remplir la parole, trois paroles, une parole, la parole) sont strictement inadmissibles pour cette acception du mot *word* (parole).

c) l'aptitude à exprimer la même pensée de plusieurs façons différentes si le message n'a pas été interprété correctement ou tout simplement suivant les circonstances, le caractère stylistique et le contexte général dans lequel se déroule un acte de communication : *The nestlings died as a result of the cold wave, The nestlings perished as a result of the cold wave, The nestlings died due to the cold wave, The cold wave killed the nestlings, The cold wave caused the death of the nestlings, The cold wave led to the death of the nestlings, The death of the nestlings was a consequence of the cold wave, The death of the nestlings was due to the cold wave, The death of the nestlings was a result of the cold wave, The nestlings died as a result of a sudden drastic*

drop in temperature, The nestlings perished as a result of a sudden drastic drop in temperature, The nestlings died due to a sudden drastic drop in temperature, etc.

Toutes ces aptitudes peuvent être décrites dans le cadre de la théorie « Sens-Texte » en termes de décomposition sémantique et de fonctions lexicales. Ces dernières permettent à la fois de formuler les règles de la cooccurrence des lexèmes et celles du paraphrasage.

2. Fonctions lexicales

Une FL, selon Mel'čuk, est un **sens** suffisamment abstrait et général qui peut être exprimé par un grand nombre de lexèmes différents ; le choix d'un lexème concret est chaque fois déterminé par le mot-clé (**argument**) auquel ce **sens** abstrait est associé. En d'autres mots, les FL sont des **sens**, dont l'expression est soumise à toutes sortes de restrictions lexicales. Prenons quelques exemples avec la FL bien connue MAGN (intensificateur) qui signifie « le haut degré de ce qui est exprimé par le lexème-argument ».

Fonction	**Argument**	**Valeur**
MAGN	*disease*	*grave*
MAGN	*contrast*	*sharp*
MAGN	*control*	*strict*
MAGN	*sleep (verb)*	*soundly*
MAGN	*know*	*firmly*

Un autre exemple des FL sont les verbes-supports appartenant à la famille OPER-FUNC, à savoir : $OPER_1$, $OPER_2$, $FUNC_1$, $FUNC_2$, $LABOR_{1-2}$, $LABOR_{1-2}$. Ce sont des verbes sémantiquement « vides » qui en fait reprennent une partie du sens exprimé par le mot-clé (argument). Leur apport sémantique à la signification de la phrase est quasiment nul. Notons que du point de vue syntaxique ces verbes sont des conversifs les uns par rapport aux autres.

$OPER_1$ prend le premier actant de la situation exprimée par le mot-clé en tant que sujet grammatical et il prend le mot-clé en tant que complément d'objet direct :

Fonction	**Argument**	**Valeur**
$OPER_1$	*resistance*	*put up (resistance)*
$OPER_1$	*control*	*exercise (control)*
$OPER_1$	*operation (surgical)*	*make (an operation)*
$OPER_1$	*fear*	*feel (fear)*
$OPER_1$	*proposal*	*make (a proposal)*
$OPER_1$	*sound*	*utter (a sound)*
$OPER_1$	*care*	*take (care) of (smb)*

$OPER_2$ prend le deuxième actant de la situation exprimée par le mot-clé en tant que sujet grammatical et il prend le mot-clé en tant que complément d'objet direct :

$OPER_2$	*resistance*	*meet (resistance)*
$OPER_2$	*control*	*be under (control)*
$OPER_2$	*operation*	*undergo (an operation)*

$FUNC_1$ prend le mot-clé en tant que sujet grammatical et il prend le premier actant de la situation exprimée par le mot-clé en tant que complément d'objet direct :

$FUNC_1$	*fear*	*possesses (smb)*
$FUNC_1$	*proposal*	*comes from (smb)*
$FUNC_1$	*sound*	*escapes (smb)*

$LABOR_{1-2}$ prend le premier actant de la situation exprimée par le mot-clé en tant que sujet grammatical, il prend le deuxième actant en tant que complément d'objet direct et le mot-clé en tant que complément d'objet indirect :

$LABOR_{1-2}$	*control*	*keep (smth) under (control)*
$LABOR_{1-2}$	*criticism*	*subject (smb) to (criticism)*
$LABOR_{1-2}$	*care*	*have (smb) in (one's care)*

Les quatre traits spécifiques des FL indiqués ci-dessous les rendent particulièrement attrayantes pour l'étude des langues : a) universalité ; b) adaptation au paraphrasage ; c) spécificité intralinguistique (idiomatisme) ; d) spécificité interlinguistique (idiomatisme).

a) Dans le cadre de la théorie « Sens-Texte » on définit environ 50 FL élémentaires, que l'on suppose être universelles pour toutes les langues et qui sont, par conséquent, décrites au moyen d'un métalangage formel standard. Ce métalangage qui s'appuie sur les assises théoriques très solides permet d'établir des équivalences entre les expressions de n'importe quelles langues sans jamais recourir à la traduction « mot à mot » – péché commun d'un grand nombre de dictionnaires. Nous avons déjà cité plus haut quelques exemples des FL MAGN en anglais, il serait intéressant de voir maintenant leurs équivalents russes :

Fonction	**Argument**	**Valeur**
MAGN	*bolezn' 'disease'*	*tjazhelaja*, lit. *'heavy'*
MAGN	*kontrast 'contrast'*	*rezkij*, lit. *'cutting'*
MAGN	*kontrol' 'control'*	*strogij*, lit. *'strict'*
MAGN	*spat' 'to sleep'*	*krepko*, lit. *'firmly'*
MAGN	*znat' 'to know'*	*tverdo*, lit. *'hard'*

b) La bonne adaptation des FL au système de paraphrasage peut être illustrée sur l'exemple de la famille OPER-FUNC. Comme nous l'avons déjà signalé plus haut, l'apport sémantique des FL, faisant partie de cette famille, à la signification du mot-clé et à toute la phrase est minime. Aussi constituent-elles un kit linguistique idéal permettant de formuler quelques règles de paraphrasage très générales et d'établir les relations de synonymie entre les énonciations. Notamment, chaque verbe « significatif » X peut être paraphrasé moyennant une expression constituée par un verbe support, vide de sens, appartenant à la famille OPER-FUNC et un substantif – nom d'action, d'état ou de processus – dérivé du verbe X.

$X = OPER_1 + S_0(X)$
$X = OPER_2 + S_0(X)$
$X = LABOR_{1-2} + S_0(X)$
$X = FUNC_1 + S_0(X)$ etc.

Les équations ci-dessus permettent de construire les équivalences suivantes : $OPER_1 + S_0(X) = OPER_2 + S_0(X) = LABOR_{1-2} + S_0(X) = FUNC_1 + S_0(X)$ etc. Cf. :

The president controls (X) *the situation* =
The president has ($OPER_1$) *control* ($S_0(X)$) *of the situation* =
The president keeps ($LABOR_{1-2}$) *the situation under control* ;
The surgeon operated (X) *on the wounded* =
The surgeon made ($OPER_1$) *an operation* ($S_0(X)$) *on the wounded* =
The wounded underwent ($OPER_2$) *an operation (by the surgeon)* ;
He was afraid (X) =
He felt ($OPER_1$) *fear* ($S_0(X)$) =
Fear possessed ($FUNC_1$) *him* ;
He didn't utter ($OPER_1$) *a sound* =
No sound escaped ($FUNC_1$) *him.*

L'aptitude à trouver ces paraphrases fait partie des facultés linguistiques qui assurent la maîtrise parfaite d'une langue, qu'il s'agisse d'une langue étrangère ou maternelle.

c) La notion de spécificité intralinguistique est assez transparente et ne nécessite pas de longues explications, il suffit juste de regarder quelques exemples des FL cités ci-dessus pour se rendre à l'évidence qu'il n'existe aucune corrélation sémantique directe entre la valeur d'une FL et le sens du mot-clé. En effet, il serait très difficile d'expliquer pourquoi *orders are issued*, alors que *demands are made* ; ou encore pourquoi nous disons *sleep soundly*, mais *know firmly*, et pourquoi le contraire est impossible : **sleep firmly*, **know soundly* – malgré le fait que du point de vue sémantique les deux adverbes ont une signification très proche. Ce phénomène devient très spectaculaire lorsqu'on compare deux synonymes, pour lesquels les restrictions de cooccurrence lexicale sont différentes. Ainsi, il n'existe aucune raison sémantique pour que le substantif *word* lorsqu'il signifie 'promesse' ne puisse être employé avec les verbes *to make* et *to fulfil* (*to *make a word*, **to fulfil a word*), – qui par ailleurs s'emploient facilement avec le substantif *promise*. Ce qui plus est, deux substantifs, dont les acceptions principales sont synonymiques, par exemple, *desire* et *wish*, sont parfois soumis à des restrictions différentes ; cf. *strong/keen/intense/fervent/ardent/ overwhelming desire*, mais seulement *strong/fervent/ardent wish*. Par conséquent, une substitution synonymique libre est impossible pour ces deux mots. Ainsi, toute personne qui prétend parler couramment une langue, doit pour chaque mot-clé apprendre tout bêtement par cœur le lexème qui signifie le haut degré de ce qui est exprimé par ce mot-clé, puisqu'il n'existe aucun critère sémantique qui lui permette de choisir le mot approprié dans chaque cas concret.

d) Compte tenu de la très grande variété d'expressions d'une fonction lexicale au sein d'une langue donnée, on peut présumer que la spécificité interlinguistique des FL est encore plus prononcée. Pour ne pas citer de nouveaux exemples, nous renvoyons à ceux de la FL MAGN que nous avons considérés plus haut dans le paragraphe (a) en comparant les substantifs russes à leurs équivalents anglais : *bolezn'/disease, kontrast/ contrast, kontrol'/control, spat'/to sleep, znat'/to know*. Il existe cependant un autre aspect – beaucoup plus fondamental – de la spécificité interlinguistique des FL : les langues sont différentes non seulement parce qu'elles expriment la même idée par des mots différents (comme nous l'avons vu dans les exemples précédents), mais aussi et surtout

parce que la fréquence de l'expression de cette idée varie d'une langue à l'autre. Cette dernière différence est profondément ancrée dans la typologie d'une langue donnée et dans sa base conceptuelle inhérente (la soi-disant « vision naïve de l'univers »). Ainsi, la langue russe qui tend à masquer le vrai agent d'une action et à personnifier les états physiques et mentaux en leur prêtant une volonté autonome et indépendante (d'où la profusion des constructions impersonnelles et indéfinies en russe), utilise beaucoup plus que l'anglais les verbes supports qui prennent le mot-clé (argument d'une FL) en tant que leur sujet grammatical. Alors que l'anglais ou le français, avec leur rationalisme et leur tendance à mettre en relief l'agent d'une action, évitent les constructions impersonnelles et essayent de les remplacer par les constructions agentives. Cf. :

U nego ['*he*', complément d'objet] *byli opasenija*
['*apprehensions*', sujet grammatical] *po aetomu povodu* -
He [sujet grammatical] *had certain apprehensions*
[complément d'objet] *on this account*.

où les mots russes *on, opasenija* et leurs équivalents anglais *he* et *apprehensions* jouent les rôles syntaxiques opposés. Cf. aussi :

Toska glozhet <snedaet, tochit> ego [complément d'objet] -
He [sujet grammatical] *is consumed with melancholy,*
Radost' vladeet im [complément d'objet] -
He [sujet grammatical] *is full of joy,*
Ljubopytstvo razbiraet ego [complément d'objet] -
He [sujet grammatical] *cannot suppress his curiosity,*
Revnost' muchit ego [complément d'objet] -
He [sujet grammatical] *is a martyr to jealousy,*
U nego [complément d'objet] *na ume chto-to drugoe* -
He [sujet grammatical] *has something else on his mind etc.*

Cette construction syntaxique s'applique en russe non seulement aux substantifs qui désignent les émotions et les états mentaux, mais aussi aux noms d'états physiques, comme par exemple *kashel'* 'toux', *golod* 'faim', *zhazhda* 'soif', etc. :

Ego [complément d'objet] *bil <donimal> kashel'* -
He [sujet grammatical] *had a bad fit of coughing,*
Ego [complément d'objet] *muchit golod <zhazhda>* -
He's [sujet grammatical] *going through the tortures of hunger <thirst>.*

Cf. aussi :
Èta melodija do six por zvuchit u menja [complément d'objet] *v ushax* -
I [sujet grammatical] *can still hear this melody,*
Protivnyj vkus ryby do six por u menja [complément d'objet] *vo rtu* -
I [sujet grammatical] *can still feel the foul taste of fish in my mouth,* etc.

Ce type de divergences entre les langues peut être explicité facilement en termes de FL. Ainsi, en décrivant la façon d'exprimer les états physiques, mentaux et émotionnels dans les différentes langues, on pourra dire que l'anglais préfère les constructions à $OPER_1$ là où le russe emploie les constructions à $FUNC_1$.

Dans le cadre de la théorie « Sens-Texte », on distingue deux grandes catégories de FL : les FL (dites **paradigmatiques**) qui s'emploient *à la place* du mot-clé et les FL **syntagmatiques** qui s'emploient *avec* le mot-clé (par exemple, la FL MAGN).

Parmi les FL paradigmatiques on retrouve : SYN (synonymes), ANTI (antonymes), CONV (conversifs, comme *to buy* et *to sell*), GENER (mot générique ou hyperonyme), ainsi que les dérivés syntaxiques d'un mot-clé : cf. S_0(*to teach*) = *teaching*, S_1(*to teach*) = *teacher* ('celui qui enseigne'), S_2(*to teach*) = *subject* ('ce qu'on enseigne'), S_3(*to teach*) = *pupil* ('celui à qui on enseigne') ; S_0(*to sell*) = *sale*, S_1(*to sell*) = *seller*, S_2(*to sell*) = *article*, S_3(*to sell*) = *buyer*, S_4(*to sell*) = *cost*, etc. Ces FL sont largement utilisées dans certains types de paraphrases, cf. *He teaches me* = *He is my teacher* = *I am his pupil* (basées sur les règles universelles ci-dessous. X = copula + S_1(X) = copula + S_2(X), etc.).

Les LF syntagmatiques comprennent avant tout les verbes supports $OPER_1$, $OPER_2$, $LABOR_{1-2}$ etc. ; elles participent également à un certain nombre de paraphrases (voir les exemples ci-dessus).

Par ailleurs, les FL se subdivisent en FL **simples ou élémentaires** (voir les exemples ci-dessus) et en FL **complexes**, ces dernières étant généralement construites par superposition de deux ou plusieurs FL simples. Voici quelques exemples de FL complexes (dont les noms sont suffisamment parlants et ne demandent pas d'explications supplémentaires) : $INCEPOPER_1$, $FINOPER_1$, $INCEPFUNC_1$, $FINFUNC_1$, etc.

Le nombre total des principales FL simples et complexes, dont on aura besoin pour formuler les règles du choix lexical collocationnel et du paraphrasage, s'élève à environ 80. La liste des FL a été revue et adaptée aux besoins des jeux linguistiques. Pour chacune des FL faisant partie de cette liste nous avons fourni une définition en anglais et en russe et un ensemble de mots (plusieurs dizaines pour certaines fonctions) qui servent à illustrer cette fonction et qui constituent la « matière linguistique » pour le jeu. Par rapport à la version initiale de la théorie STM, les définitions des FL ont été systématisées et standardisées.

3. Décomposition sémantique (définition)

Une entrée du DEC russe ou anglais contient les informations suivantes : 1) une vedette ou nom du lexème (par lexème on entend un mot dans une de ses acceptions lexicales) ; 2) une définition du lexème (d'une acception lexicale) ; 3) une partie du discours ; 4) un équivalent du lexème dans la langue cible ; 5) les principales FL paradigmatiques pertinentes pour ce lexème ; 6) les principales FL syntagmatiques pertinentes pour ce lexème.

En dehors des FL, l'information la plus importante d'une entrée du DEC est la définition. Les définitions, quoique un peu simplifiées, sont formulées dans un métalangage sémantique et s'appuient sur quelques principes généraux posés par l'auteur du manuel dans ses publications récentes.

Le métalangage sémantique n'est en fait qu'un sous-langage d'une langue naturelle donnée. Ceci est vrai tant pour son lexique que pour sa grammaire. Ce mé-

talangage comprend des mots et des constructions syntaxiques relativement simples. Chaque élément du métalangage sémantique doit répondre aux critères de la non-synonymie (idéalement, à chaque sens doit correspondre un seul mot) et de la non-homonymie (idéalement, à chaque mot nous devons attribuer un seul sens). La synonymie est autorisée dans deux cas bien précis : lorsqu'il s'agit d'une transposition syntaxique : *to make/making* et dans le cas des unités sémantiques élémentaires, cf. *to want* et *good*, ou *to know* et *true*, bien qu'il y ait un très grand recoupement sémantique entre les deux paires de mots ci-dessus.

Même ce descriptif très sommaire montre que notre métalangage sémantique n'est pas universel, mais qu'il est spécifique pour une langue donnée. Ceci n'exclut pas la possibilité de l'adapter à une autre langue à condition qu'il s'agisse de langues assez proches du point de vue culturel, par exemple le russe, l'anglais, le français et l'allemand. Dans la plupart des cas une simple traduction mot à mot du lexique du métalangage suffit.

Le noyau du métalexique est constitué par les soi-disant **unités sémantiques élémentaires** – mots qui ne peuvent pas être décomposés en éléments plus simples dans une langue donnée sans qu'il y ait un cercle vicieux. Ceci dit, une unité sémantique élémentaire ne doit pas nécessairement correspondre au plus simple sens possible. Le seul critère auquel elle doit répondre est le suivant : chaque fois que l'on veut discerner un sens encore plus simple dans une unité sémantique élémentaire, il faut trouver un mot approprié dans la langue en question pour exprimer ce sens. Il est évident, par exemple, que les verbes *to want, to wish* et *to desire* du point de vue sémantique ont une partie commune importante. Mais cette partie commune n'est en fait qu'une pure abstraction, que l'on n'arrive pas à verbaliser en anglais. En réalité, le verbe *to want* ne peut pas prétendre à exprimer un sens plus élémentaire, puisqu'en plus du noyau sémantique commun pour tous ces verbes il contient aussi l'idée de la 'nécessité', qui est totalement étrangère aux verbes *wish* et *desire* ; à son tour le verbe *wish* peut exprimer l'idée de la 'futilité' des désirs du sujet, alors que le verbe *desire* rajoute l'idée de 'volonté' et de 'résolution'.

La liste des unités sémantiques élémentaires en anglais contient les éléments, tels que : *time, space, world, object, person, property, part, state, process, eye, ear,* ainsi que *good, true, all, one, first, to want, to think, to know, to feel, to say, to do, to happen, to be, to perceive, to cause, can, not, or, and that, what, which* etc. En dehors des unités sémantiques élémentaires, le métalangage contient aussi quelques éléments sémantiques intermédiaires, comme : *obligatory = such that it is impossible not to do it ; impossible = not possible ; possible = such that can happen or be done.*

Passons maintenant aux définitions à proprement parler. Elles doivent répondre aux critères suivants :

1) absence de cercles vicieux ;
2) exhaustivité et non-redondance : la définition d'un lexème doit contenir tous ceux et seulement ceux des éléments sémantiques qui constituent son sens ;
3) réductibilité : il faut que toute définition puisse être décomposée en unités sémantiques élémentaires, directement, ou en passant par des stades intermédiaires ;
4) systématicité : l'ensemble des définitions doit être construit de sorte que l'on puisse identifier facilement les liens sémantiques systématiques qui existent entre les lexèmes d'une langue.

Pour illustrer les concepts généraux exposés ci-dessus, nous allons proposer les définitions d'un groupe de lexèmes (dits « temporaux ») qui présentent des traits sémantiques communs : *autumn, dark, day$_1$, day$_2$, evening, hour, January, light, minute, Monday, month$_1$, month$_2$, morning, night, period, season, second, to see, spring, summer, sun, sunrise, sunset, today, tomorrow, week, winter, year, yesterday.*

Autumn = the season of the year between summer and winter when it starts to be getting colder.
Dark = such as has little or no light.
Day$_1$ = the part of time between two consecutive sunrises.
Day$_2$ = the lightest part of day$_1$ which follows morning, precedes evening and ends between 16 and 17 o'clock.
Evening = the part of day$_1$ when it starts to be getting dark, which includes sunset and ends between 23 and 0 o'clock.
Hour = the 24-th part of day$_1$.
January = the first month of the year.
Light = that property of the world which makes it possible to see objects.
Minute = the sixtieth part of an hour.
Monday = the first working day of the week.
Month$_1$ = one of the twelve parts into which a year is divided.
Month$_2$ = any period of approximately 30 days$_1$.
Morning = the part of day$_1$ when it starts to be getting light which includes sunrise and ends between 11 and 12 o'clock.
Night = the darkest part of day$_1$ which follows evening, precedes morning and ends between 4 and 5 o'clock.
Period = any part of time.
Season = one of the four large parts into which a year is divided on the basis of the natural cycle considerations.
Second = the sixtieth part of a minute.
See = to perceive with eyes.
Spring = the season of the year between winter and summer when it starts to be getting warmer.
Summer = the warmest season of the year with the longest days$_2$ and shortest nights.
Sun = the brightest object in the sky which gives light and warmth.
Sunrise = the process of the sun appearing on the horizon and the initial stages of its going up in the sky, or the time during which this process takes place.
Sunset = the process of the sun disappearing beyond the horizon and the final stages of its going down in the sky, or the time during which this process takes place.
Today = that day$_1$ through which the speaker is living at the time of speech.
Tomorrow = the day$_1$ immediately after today.
Week = a period of seven days$_1$ singled out on the basis of human activity cycle considerations.
Winter = the coldest season of the year with the shortest days$_2$ and longest nights.
Year = the part of time equal to 365 days$_1$, in which the earth makes one full circle around the sun.
Yesterday = the day$_1$ immediately before today.

4. Jeux linguistiques

Le manuel permet de jouer à quatre jeux linguistiques différents basés sur les définitions des lexèmes et sur les FL et qui consistent à : 1) trouver une traduction pour le lexème proposé par l'ordinateur ; 2) indiquer les valeurs de toutes les FL affichées par l'ordinateur pour un lexème choisi par le joueur ; 3) indiquer les valeurs d'une FL choisie par le joueur pour tous les lexèmes affichés par l'ordinateur ; 4) trouver le nom du lexème qui correspond à la définition affichée par l'ordinateur.

Dans sa version initiale le manuel supposait que tous les jeux seront joués dans le cadre d'une même langue. Mais maintenant chacun des jeux énumérés ci-dessus peut aussi être joué en mode bilingue, où la connaissance de sa langue maternelle (quelle qu'elle soit) aide le joueur à trouver une réponse correcte. Ceci est possible dans la mesure où l'on arrive à garder l'isomorphisme des DEC.

Le manuel permet également de vérifier la qualité des connaissances linguistiques d'un joueur. À chaque étape du jeu on affiche le nombre de points que le joueur a gagnés par sa dernière réponse et le score total du jeu.

Nous avons prévu de doter le manuel d'un système de dialogue moderne (actuellement notre programmeur travaille à la mise au point de ce menu) basé sur une présentation idéographique qui permettra aux joueurs de choisir le domaine lexico-sémantique qui les intéresse, le type et le mode du jeu.

Nous travaillons également à la création d'un logiciel qui doit permettre l'extension et la mise à jour des DEC et des autres ressources linguistiques du manuel.

Par la suite nous envisageons également d'introduire de nouveaux jeux linguistiques, dont un jeu de paraphrasage qui consiste à trouver le maximum de paraphrases possibles pour une proposition donnée construite à partir des unités lexicales présentes dans les DEC.

L'hypertexte *Hyperbase*

Étienne Brunet

Institut national de la langue française, UPR 6861 (CNRS), Nice, France

On a quelque scrupule à présenter un produit qui n'est plus tout jeune ni tout à fait inconnu. Mais sa notoriété n'est pas suffisante pour autoriser la prétérition et son âge – six ans déjà – mesure en réalité le destin d'une famille où plusieurs générations sont enveloppées. Tout logiciel qu'on a l'imprudence de livrer au public est en effet un boulet qu'on traîne pendant des années, avec la nécessité de le renouveler sans cesse, pour le corriger, l'améliorer, le compléter et surtout l'adapter aux changements rapides des machines et des systèmes. Quelle différence avec le livre, bon ou mauvais, qui s'envole au moment de la publication et dont l'auteur se libère d'un coup, en songeant au suivant. Quand on met un logiciel sur le marché on ne peut jamais dire *alea jacta est*. Le lancement d'un tel produit est au mieux celui d'un cerf-volant captif dont il faut gouverner les sautes en surveillant le vent, en sorte que l'auteur est plus captif encore.

Curieusement les auteurs de logiciels ont pris l'habitude, chaque fois qu'ils publient une version nouvelle, de faire l'historique de leur produit, en dressant le catalogue de leurs fautes, de leurs corrections, de leurs variations. Quel soin scrupuleux pour noter les erreurs et les repentirs, alors que les écrivains partagent généralement le souci inverse, qui est d'effacer les ratures et de laisser croire au premier jet lumineux de l'inspiration. Et ce blanchissement est facilité de nos jours par les traitements de texte qui anéantissent les brouillons et toutes les traces de la transpiration. Cette discrétion orgueilleuse étant réservée à l'écrivain, on s'en tiendra donc à la pratique des auteurs de logiciels dont la modestie excuse l'impudeur.

1. La première version d'*Hyperbase*

La tradition des concordances est ancienne et l'idée d'en confier la réalisation à l'ordinateur est venue très tôt, dès les années 60. Les premières réalisations, à Besançon (avec Quemada), à Liège (avec Delatte), à Gallarate (avec Busa), datent de cette époque héroïque. Dix ans plus tard, les gros systèmes offraient à peu près partout dans le monde de tels services documentaires : à Nancy, à Paris, à Pise, à Oxford, à Göte-

borg, à Montréal et dans beaucoup d'universités américaines. Et dès 1970 nous proposions une chaîne de programmes pour le traitement documentaire et statistique des données textuelles[1]. À cette époque déjà le *TLF* avait mené à son terme l'indexation du corpus des XIX^e^ et XX^e^ siècles, soit plus de 70 millions de mots. Et disposant des données dérivées de ce traitement (cela s'appelait les fichiers-répertoires), nous avons créé une panoplie d'outils spécialisés pour l'étude quantitative de cette matière textuelle à demi traitée, au besoin en faisant le chemin inverse et en reconstituant le texte à partir des index[2]. Dix ans plus tard, au cours des années 80, le paysage technologique change radicalement avec l'avènement de la micro-informatique et l'extension des réseaux. À l'Institut national de la langue française se constitue la base de données *Frantext*, qu'on peut interroger par le réseau *Transpac* et qui offre à la communauté scientifique un immense champ de recherche (près de 3 000 textes complets et quelque 160 millions de mots), en même temps qu'un produit dérivé, *Discotext 1*, présente une large part de cette base (500 titres) sur CD-ROM, au standard PC.

De notre côté nous nous étions lancé dans un projet, complémentaire des précédents, et destiné aux utilisateurs (nombreux chez les littéraires) du standard Apple. Et un contrat de développement, signé avec la compagnie Apple-France, prévoyait une gamme de produits diversifiés. Le premier prototype d'*Hyperbase* était destiné, comme *Discotext*, aux grands corpus. Conçue en 1989 pour une manifestation du Bicentenaire à Beaubourg, cette version première du logiciel a mis à la disposition du public du Centre Pompidou un ensemble de textes de la Révolution, issu du *TLF* et représentant 30 millions de caractères. Il n'est guère utile d'en décrire le détail, puisque deux publications sont explicites là-dessus[3]. Il suffit de représenter, dans la figure 1, le menu qui accueille l'utilisateur et qui lui est offert systématiquement dès qu'une action est accomplie. Deux voies principales s'ouvrent à la sélection : à droite on s'engage dans la recherche documentaire ; à gauche on s'oriente vers les traitements statistiques.

1. « Programmes linguistiques, série 1 », *CUMFID*, n° 2, 1970, 175 p. Cette chaîne de programmes intégrés a servi en particulier à l'élaboration des *Index de J. J. Rousseau* en 26 volumes, publiés aux Éditions Slatkine, Genève.

2. Plusieurs monographies sont nées de cette exploitation intensive du gisement de Nancy, sur Proust, Zola, Hugo, ouvrages publiés aux Éditions Slatkine.

3. « Hyperbase : logiciel documentaire et statistique pour l'exploitation des grands corpus », *Tools for Humanists*, Toronto, 1989, pp. 33-36.

« Computer processing and quantitative text analysis : *Hyperbase*, an interactive software for large corpora », *Data Analysis, Learning Symbolic and Numeric Knowledge*, INRIA, Nova Science Publishers, New York, Budapest, 1989, pp. 207-214.

« What do the tables say ? What do the figures say ? », Colloque ALLC de Toronto *The Dynamic Text*, juin 1989, in *Literary and Linguistic Computing*, Oxford, 1989, pp. 70-82.

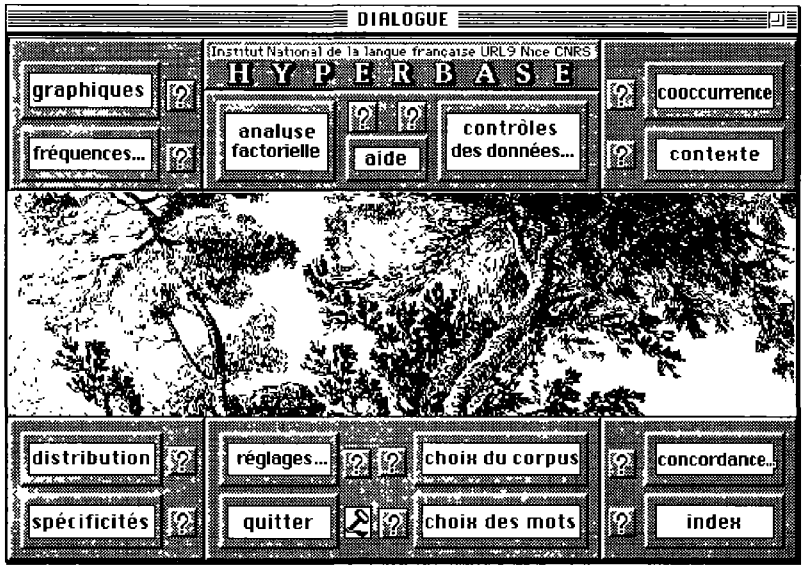

FIGURE 1 : *Hyperbase* première version. Page d'accueil.

Le choix du traitement s'accompagne d'autres sélections qui s'exercent parmi les textes et parmi les mots et que proposent les deux écrans ci-dessous (figures 2 et 3). Ces fonctions de sélection et d'exclusion se justifient dans le cas des très grands corpus qui réunissent une multitude de textes, de genres, d'époques et d'auteurs. Elles sont naturellement disponibles dans *Frantext* et *Discotext 1*. Elles permettent au chercheur de se constituer un lot de sous-corpus spécifiques et d'écarter provisoirement le reste. De même on peut à sa guise créer des listes de mots (au besoin à l'aide de filtres comme la finale, l'initiale ou la fréquence), les appliquer à des corpus différents et les retrouver d'une séance à l'autre.

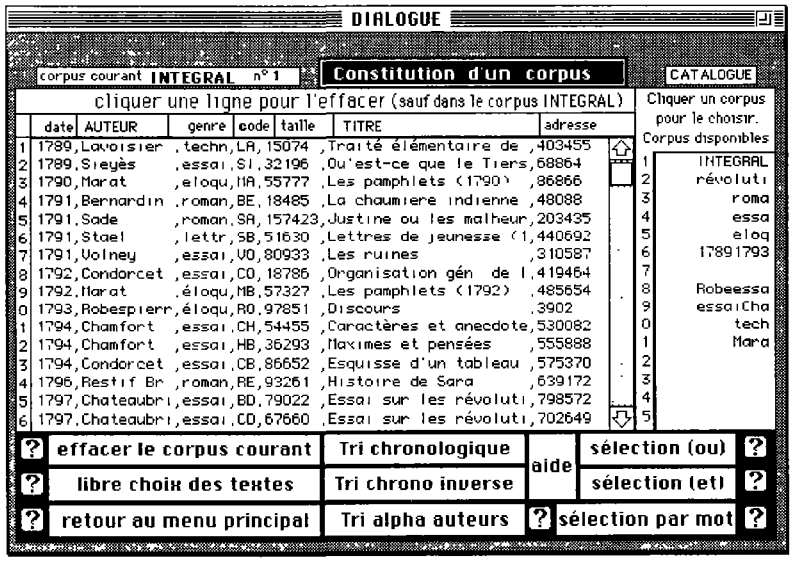

FIGURE 2 Choix du corpus.

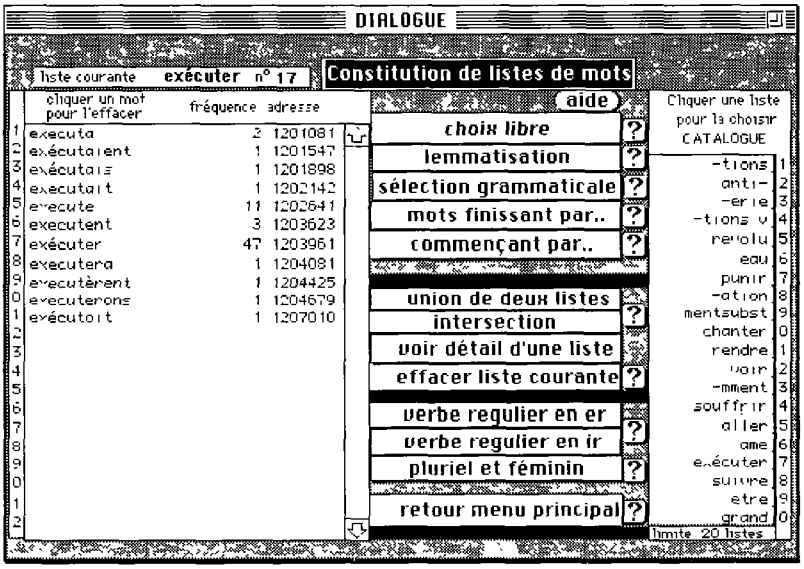

FIGURE 3 : Choix des mots

Nous ne cacherons pas notre préférence pour ce premier-né, qui nous a servi fidèlement et que nous gardons pour nos besoins personnels ou pour quelque grande entreprise, comme le CD-ROM Balzac envisagé à l'occasion du centenaire de 1999. Quand un corpus comporte une centaine de textes, il est déraisonnable de penser que les utilisateurs exploreront toujours la totalité indivisible, et peu rentable de poursuivre les recherches sur l'ensemble quand seule une fraction est jugée digne d'intérêt.

En dehors de cette liberté dans le choix de la portée, restreinte ou étendue, des traitements, la version d'origine offrait des fonctionnalités qui ont été abandonnées par la suite, à notre grand regret. Les options les plus précieuses et les plus difficiles à mettre en œuvre sont relatives à la lemmatisation et au codage grammatical. Comme on disposait d'un fichier de correspondance entre les vocables et les formes, établi à Nancy pour les besoins du *TLF*, on avait là le moyen de procéder à un regroupement des formes sous le même lemme. De plus le code grammatical emprunté à la même source permettait d'ajouter un filtre à la sélection des mots et par exemple de réunir les mots en *-tions* qui ne soient pas des formes verbales (à la 1re personne du pluriel). Enfin des données quantitatives puisées dans le fichier du *TLF* permettaient d'établir pour chaque vocable une comparaison entre les fréquences observées dans le corpus de la Révolution et dans un corpus plus vaste englobant la première moitié du XIXe siècle.

2. La version commercialisée d'*Hyperbase*

Comme le prévoyait le contrat de développement signé avec *Apple*, une nouvelle version du logiciel a été créée, qui se fonde sur des principes différents et dont la logique est celle des logiciels commerciaux, qui sont vides de données (comme les traite-

ments de texte, les tableurs, etc.), mais riches d'outils variés, conçus pour aider l'utilisateur à traiter son bien propre. Le programme a certes un jeu de données provisoires, ce qui facilite l'apprentissage. Mais la base peut être vidée de son contenu et recevoir d'autres données (sous forme de texte *ASCII*). Programmes de préparation et d'exploitation sont fournis conjointement dans le même produit.

Hyperbase dans cette version à tout faire vise à la généralité et à la simplicité. Le produit est conçu pour s'adapter immédiatement aux données de l'utilisateur et réaliser l'indexation dans un temps acceptable et sans manipulation excessive. Ces contraintes ont empêché de fournir les outils spécialisés de la lemmatisation, laquelle est nécessairement propre à chaque langue et ne peut jamais être complètement automatique. Comme cette version indifférenciée du logiciel devait s'appliquer à tous les textes qui utilisent un alphabet latin et donc à la plupart des langues occidentales[4], il était difficile de fournir pour toutes ces langues un dictionnaire-machine doté non seulement des codes grammaticaux indispensables mais aussi d'indications de fréquence[5]. Si donc la lemmatisation a été sacrifiée, faute de pouvoir être universelle, les deux objectifs antérieurs ont été maintenus qui orientent l'exploitation vers la recherche documentaire et la statistique.

a - Le programme d'exploitation répond, par les méthodes de l'hypertexte, aux besoins classiques du traitement automatique des textes : concordances de type *kwic* (avec tri des expansions droite ou gauche du contexte), index sélectifs ou systématiques, dictionnaires des fréquences, sélection de contextes larges, cooccurrences, filtrage et masquage des mots et constitution de listes, recherche des parties de mots (début, fin ou chaîne quelconque) ou des groupes de mots, limitation ou extension du corpus de travail, etc.

Nous renvoyons le lecteur à des publications antérieures où l'on trouvera une description plus détaillée de ces fonctions documentaires[6]. Un mode d'emploi d'une centaine de pages, plus explicite encore, accompagne le logiciel et guide l'utilisateur. Celui-ci a encore à sa disposition une aide en ligne plus concentrée, et, qui plus est, une page d'explication, immédiatement disponible, pour chaque fonction. Si le symbolisme d'un bouton n'est pas compris tout de suite et si son nom laisse perplexe, on peut par précaution faire apparaître sur l'écran les instructions précises avant de déclencher l'action correspondante.

On se bornera à titre d'exemple à la fonction *contexte* dont on montrera les options (figure 4) et le résultat (figure 5). Le corpus exploré est ici celui de Julien Gracq, soit l'ensemble de son œuvre (17 titres).

4 Une version particulière a été adaptée au grec moderne La même opération pourrait être envisagée pour le cyrillique

5. Un seul dictionnaire est fourni qui est limité au français et aux 10 000 formes les plus fréquentes.

6 « Un hypertexte statistique · HYPERBASE », *JADT 1993*, TELECOM, Paris, 1994, pp 1-16
« Hyperbase, synopsis », *Traitements informatisés de corpus textuels*, Didier Érudition, 1994, pp 169-184

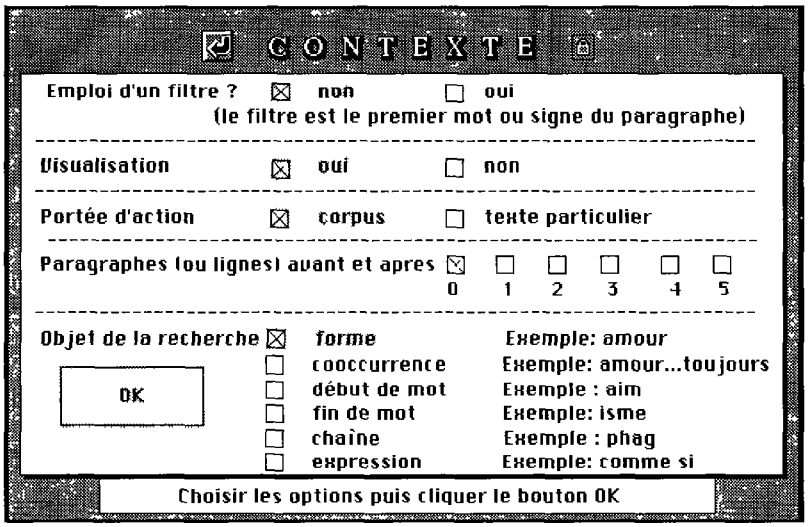

FIGURE 4 Dialogue proposé par la fonction *contexte*.

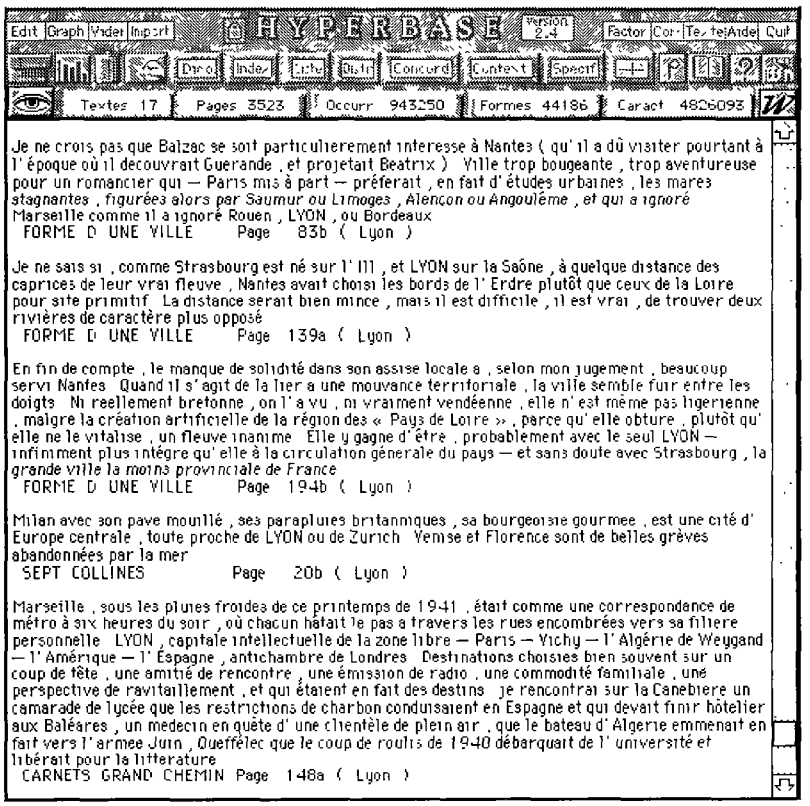

FIGURE 5 : Résultat de la fonction *contexte* dans l'œuvre de Gracq
(8 occurrences de la ville de *Lyon*, extrait).

b - *Hyperbase* se distingue des hypertextes similaires par l'orientation statistique donnée au produit. D'une part, s'il s'agit d'un texte français, une comparaison est faite[7], sous forme d'écart réduit, avec le corpus du *Trésor de la langue française* (XIX-XX^{es}, soit 70 millions de mots). D'autre part, le corpus peut être partitionné pour permettre des comparaisons internes. *Hyperbase* restitue ainsi les mots-clés propres à chaque texte, de même qu'il dresse le profil caractéristique du corpus dans son ensemble, se détachant sur la toile de fond de l'usage littéraire de la langue depuis 1789.

De même le profil d'un mot (ou de plusieurs que l'on superpose) est dessiné par *Hyperbase*, les sous-fréquences observées, judicieusement pondérées, se transformant à volonté en histogramme (figure 6). Des tableaux peuvent être constitués qui, à partir de critères, automatiques ou non, procèdent aux regroupements de mots ou de textes. Et ainsi peut-on pallier l'absence de lemmatisation. Par exemple *Hyperbase* permet de circonscrire une catégorie grammaticale, un champ thématique, voire même le système de la ponctuation. Une fois constituées, ces listes – ce sont en réalité des tableaux à deux dimensions – peuvent être soumises aux méthodes multidimensionnelles (un programme d'analyse factorielle a été intégré à *Hyperbase*).

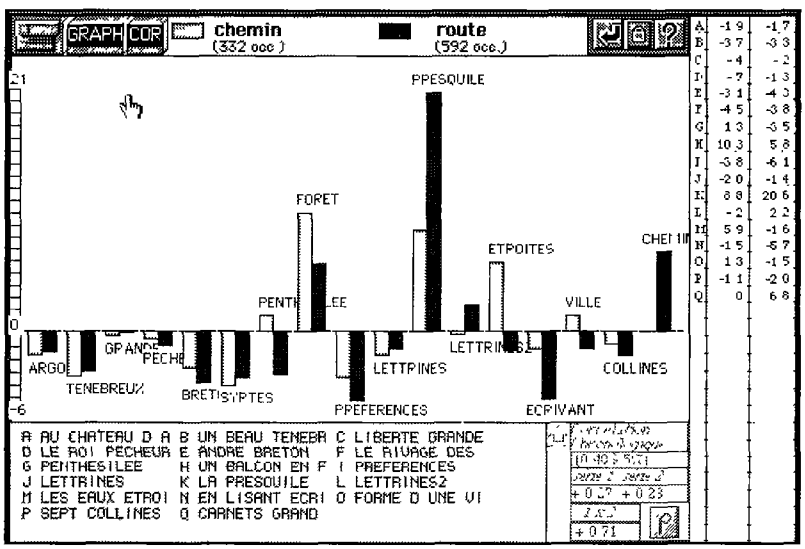

FIGURE 6 Représentation graphique de *route* et *chemin*.

D'autres calculs lexicométriques sont assurés qui permettent d'apprécier la richesse relative du vocabulaire, la distribution des classes de fréquences, l'abondance, si l'on peut dire, des mots rares (ou hapax), l'accroissement et l'évolution du vocabulaire, etc. En particulier une fonctionnalité nouvelle est apparue dernièrement (version 2.5, juillet 1995), qui mesure la distance que chaque texte établit avec tous les autres du même corpus, et qui est le rapport entre les vocables communs aux deux textes que l'on confronte et les vocables exclusifs que chacun des deux se réserve. Cela, qui peut s'appeler aussi la connexion lexicale, permet d'établir une typologie des textes à partir des similarités lexicales et principalement de leur composante sémantique[8].

7 Là où le calcul se justifie, c'est-à-dire quand la fréquence est suffisante dans le modèle (soit f = 500 dans le *TLF*).
8 La fréquence ne joue ici aucun rôle. Les effectifs sont constitués sur le seul critère présence/absence.

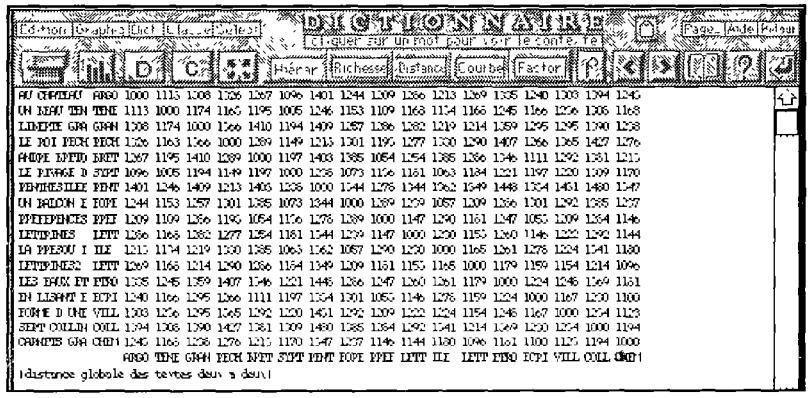

FIGURE 7 · Le tableau des distances lexicales dans l'œuvre de Gracq.

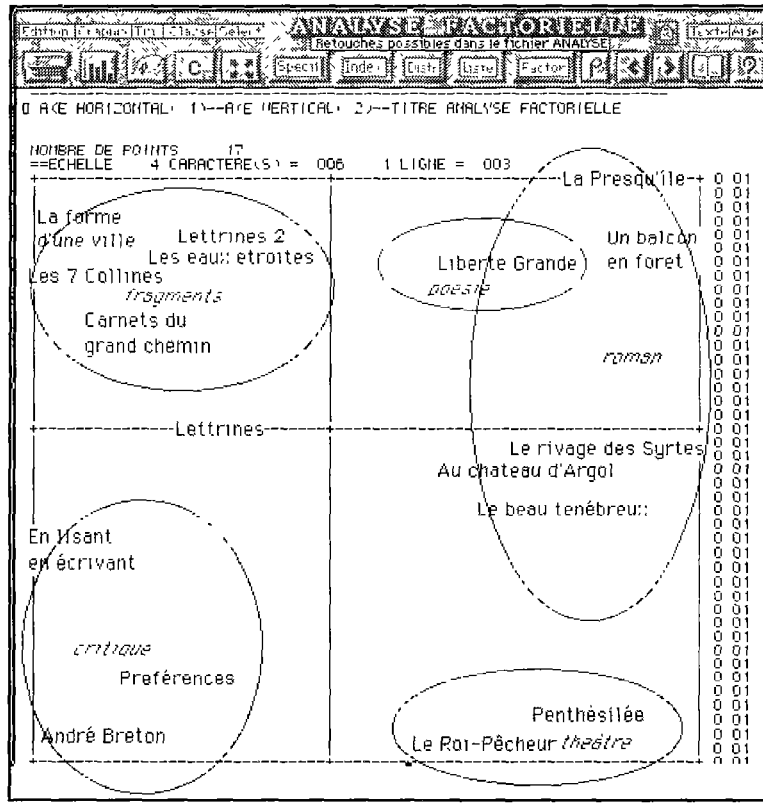

FIGURE 8 : Analyse factorielle des distances lexicales

c - Le traitement d'un nouveau texte. À la différence de beaucoup de logiciels où les fonctions sont absolument séparées des données, les unes et les autres sont mêlées dans une pile Hypercard, surtout lorsqu'il s'agit du type « standalone ». La pile originale doit donc être recopiée dès que l'on veut traiter des données nouvelles. Sous son nouveau nom elle garde ses programmes – qu'il faut conserver – et ses anciennes

données – qu'il faut évacuer. L'élimination de celles-ci se fait en sollicitant le bouton *Vider* (voir ci-dessous). Le résultat est une pile vierge, qui peut servir de modèle pour toutes les applications ultérieures. L'incorporation d'un texte est assurée par le bouton *Importer* (voir ci-dessous), qui montre la première page du fichier des données et s'enquiert des options souhaitables en affichant le dialogue de la figure 9.

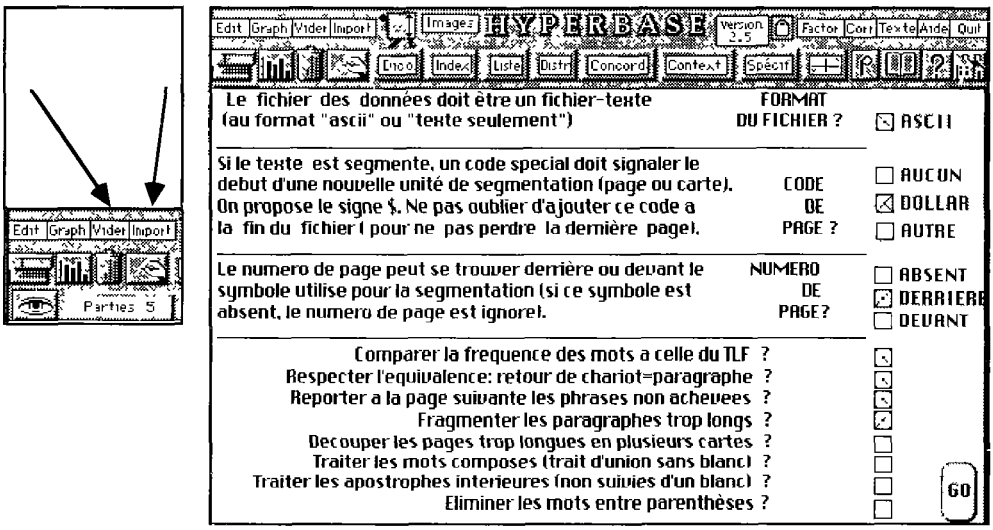

FIGURE 9 Entrée des données.

Les données textuelles doivent se trouver dans un fichier ASCII (ou « texte seulement »). On a pris en compte la plupart des alphabets européens. Aucun formatage particulier n'est obligatoire, le logiciel se chargeant de la pagination et de la partition, si elles sont absentes du fichier. En ce cas, les cartes (ou pages) ont environ 200 mots et l'ensemble du texte est découpé en dix parties de longueur voisine.

Mais il vaut mieux suivre le découpage naturel des données, s'il existe. Deux conventions doivent alors être respectées :

– les parties doivent être précédées d'une ligne où l'on indiquera le titre (en 20 caractères maximum, sans virgules ni apostrophes) en utilisant devant et derrière le symbole composite &&& (sans blanc). Veiller à bien choisir le dernier mot du titre qui sert d'abréviation lorsque la place manque, par exemple dans les graphiques, et qui doit être unique et distinctif.

– les pages sont indiquées en ajoutant une ligne (au début) et en y portant le numéro, immédiatement précédé ou suivi d'un code spécial (par exemple le symbole $; mais on peut choisir un autre code, si le symbole $ apparaît dans le texte même). Exemple :

```
&&&La vie en rose&&&
$1
texte de la page 1
$2
texte de la page 2, etc.
&&&Le travail au noir&&&
$62
texte de la page 62
$63
texte de la page 63, etc.
```

Le traitement d'un texte nouveau s'opère en trois phases : la première libère l'espace requis et transfère le texte dans la base, à raison d'une carte par page. C'est l'occasion de transcoder le texte afin d'uniformiser la présentation et en particulier de standardiser la ponctuation. En même temps, est constitué un fichier formaté qui va servir d'entrée à la phase 2. Celle-ci suit la première étape de façon automatique ou manuelle. On choisira ce dernier mode, afin de libérer le maximum de mémoire, si le fichier à traiter est de grande taille et si la machine est de faible puissance. Dans ce cas la pile est abandonnée à l'issue de la phase 1 et un double clic sur l'icône *Tripart1.6* (ci-dessous) mettra en œuvre le programme d'indexation, en lui réservant toutes les ressources de l'ordinateur.

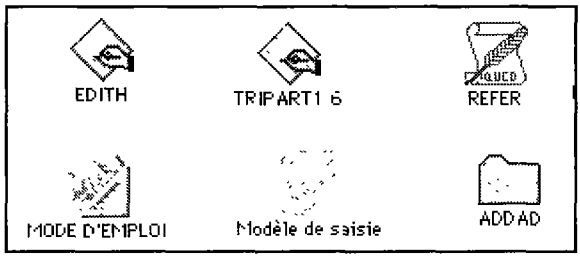

FIGURE 10 Le programme d'indexation.

Quand le tri est achevé et que les traitements subsidiaires ont pris fin, il suffit de revenir dans la pile par un double clic pour que les multiples résultats obtenus dans la phase 2 soient communiqués à la pile et définitivement enregistrés dans la phase 3. Les fichiers intermédiaires peuvent alors être détruits, la pile disposant de toutes les informations dont elle a besoin, et particulièrement du dictionnaire inverse.

Prévoir 15 minutes pour le dépouillement d'un texte de 500 pages (avec un Mac équipé d'un microprocesseur *8030*) et 20 minutes pour le tri. Après ce temps de préparation (qui comporte un transcodage, un découpage en cartes, un tri des formes, un dictionnaire des fréquences et divers tests statistiques), la pile est exploitable. Si l'on dispose d'un microprocesseur *8040* ou *PowerPC*, le temps de préparation est fortement raccourci.

d - Spécifications techniques. On ne s'appesantira pas sur les spécifications techniques. Il suffit de préciser qu'*Hyperbase* comprend un programme de préparation

(écrit en *Pascal*), un éditeur de texte (écrit en langage *C*), un programme d'analyse factorielle (écrit en *Fortran* et emprunté à *ADDAD*) et un programme d'exploitation (écrit en *Hypertalk* et complété par de nombreuses commandes externes). La configuration requise est peu exigeante et se contente d'une mémoire vive de 2 000 Ko pour son usage propre (il faut ajouter la mémoire requise par le système et celle que réclament épisodiquement les applications externes auxquelles *Hyperbase* fait appel, traitement de texte ou analyse factorielle[9]). Le disque dur est indispensable et l'écran couleur recommandé. *Hyperbase* fonctionne indifféremment sur système 6 ou 7 et sur toute la gamme *Apple*, du *Mac Plus* au *PowerMac*. Précisons que la dernière version, totalement refondue (2.5), s'est affranchie de l'environnement Hypercard. Étant du type « standalone », l'application est devenue parfaitement autonome et ne dépend plus de l'installation de l'utilisateur. Mais elle échappe aussi à toute intervention intempestive dans le code même des programmes. Les scripts sont hors d'atteinte et ni la fenêtre de commande, ni la barre de menus ne livrent accès aux intrus.

Au bout de trois ans de commercialisation, les clients sont en majorité des chercheurs du secteur public, littéraires, linguistes, historiens et sociologues mais aussi des entreprises privées, spécialisées dans la veille technologique ou les systèmes experts. À l'étranger le logiciel est plus connu au Japon, aux États-Unis, au Canada et au Brésil. Le succès est venu là où on ne l'attendait pas ; dans les instituts de sociologie ou de sondage. *Infométrie*, l'agence *Harris* et la *Sofres* se servent d'*Hyperbase*, comme le prouve l'étude du langage des principaux candidats publiée par la *Sofres*, à la veille du scrutin présidentiel de 1995.

3. La version CD-ROM d'*Hyperbase*

À partir des mêmes données relatives à l'œuvre intégrale de Julien Gracq, nous avons réalisé un premier CD-ROM, en orientant différemment le traitement. Ce CD-ROM a été présenté à Julien Gracq lui-même et communiqué à plusieurs chercheurs spécialisés dont certains œuvraient à l'édition de Gracq dans la Pléiade. C'est précisément parce que le second tome de cette édition de référence n'était pas encore paru qu'on a préféré différer la commercialisation. Il valait mieux attendre un an ou deux pour profiter du texte corrigé et pouvoir disposer de la pagination nouvelle. Au reste, Julien Gracq qui avait d'abord donné son accord, a jugé prudent d'attendre un peu afin de protéger les droits de son éditeur tant que la jurisprudence n'était pas fermement établie, en matière de *copyright*, relativement au nouveau support du CD-ROM. Actuellement des pourparlers sont en cours avec plusieurs maisons d'édition pour la diffusion du produit.

Mais entre-temps une proposition nous avait été faite qui ne soulevait pas de problèmes de cette sorte. S'agissant de Rabelais, les héritiers ne pouvaient guère avoir de prétentions non plus que les éditeurs, puisque le texte était puisé à la source. L'entreprise a démarré en juin 1994, à l'initiative de M. L. Demonet et du laboratoire *EQUIL XVI* (de l'Université de Clermont-Ferrand), qui avaient établi le texte, assuré son

9 Le seul moment où l'on a avantage à disposer d'une mémoire abondante et d'une machine rapide est celui de l'indexation, qui transforme un texte ASCII en base de données. Ce traitement exige précautions et patience mais, comme il n'a lieu qu'une fois pour chaque corpus, l'effort consenti se justifie sans peine.

transfert sur support informatique, choisi les documents et rédigé les commentaires. Restait à réunir ces matériaux sur l'étroite surface d'un CD-ROM et à en faire une base de données. En réalité, si léger que soit ce support, sa contenance est très large et dépasse de très loin le volume de la base (grosse de 120 Mo), et, en y accumulant 400 documents iconographiques, des visites guidées, une concordance complète, un index, un dictionnaire et un manuel, c'est à peine si l'on utilise la moitié (soit 314 Mo) de la surface disponible, comme le montre l'image de son contenu (figure 11). Il est vrai que la moitié restante a été consacrée au standard PC, où tous ces éléments sont accessibles, à l'exception de la base elle-même, qui est trop étroitement liée à la *Toolbox Apple* pour être portable sur un autre équipement.

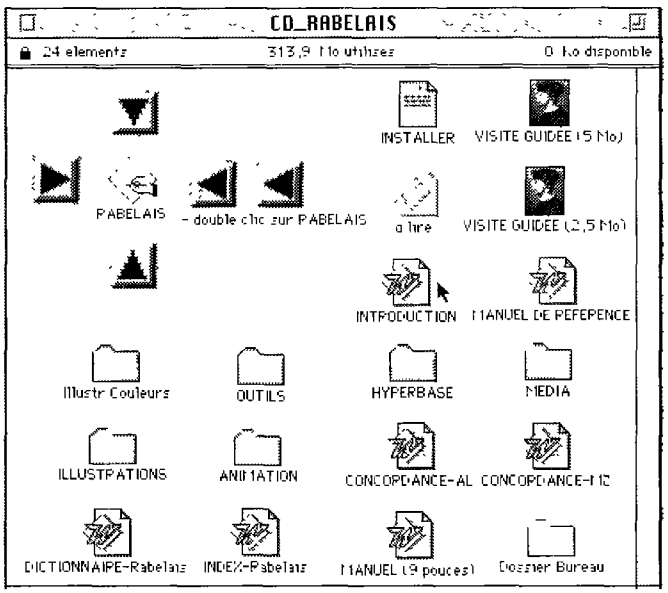

FIGURE 11 . Le contenu du CD-ROM Rabelais

Il faut souligner le caractère collectif de la réalisation, qui a bénéficié de collaborations multiples mettant en cause des organismes publics et des entreprises privées. Outre le laboratoire EQUIL XVI, à qui revient l'initiative et la conduite du projet, et le centre niçois (UPR 63861, INaLF, CNRS) qui en a assuré la réalisation technique, l'opération a reçu le concours de la Bibliothèque municipale de Lyon, de la Bibliothèque nationale de France, du Centre national du Livre et de la société APPLE-France, à quoi s'est ajoutée en dernier ressort l'industrie de l'éditeur[10]. On trouvera ci-dessous la liste des contributeurs, telle qu'elle apparaît lors de la mise en route (figure 12).

Les particularités de cette version tiennent aux spécifications du support, qui est réputé lent, tant pour l'accès que pour le débit. On a donc renversé l'ordre des priorités, en privilégiant le paramètre temps, quitte à perdre de l'espace et à redoubler l'in-

10 Éditions *Les Temps qui courent*, 118-130 bd J Jaurès, 75019 Paris

formation pour la rendre accessible à l'endroit où on la réclame, en évitant les voyages inutiles. Au demeurant le corpus est d'une taille modeste, même s'il contient dix textes qui enveloppent non seulement l'œuvre de Rabelais, de *Gargantua* au *Cinquième Livre*, mais aussi celle des devanciers et des imitateurs, soit cinq textes pararabelaisiens.

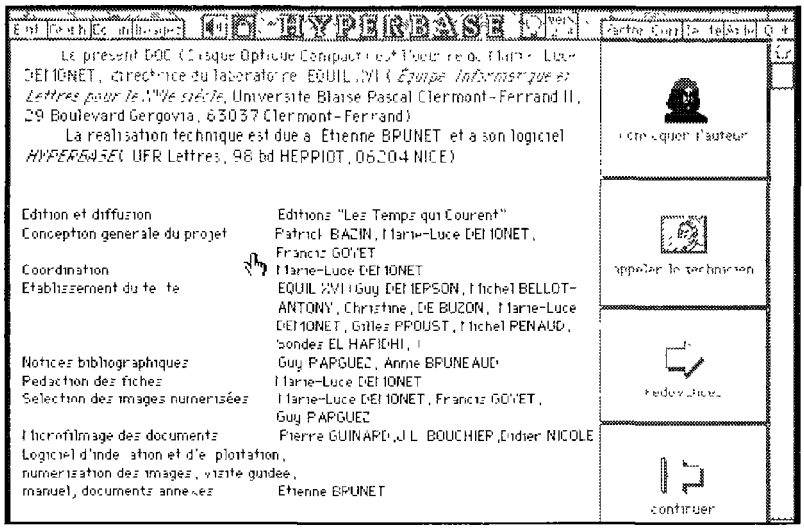

FIGURE 12 : Les contributions.

Mais la base a pris de l'embonpoint, parce qu'elle voulait être agile, si paradoxal que cela puisse paraître. On peut vérifier la vitesse du traitement en proposant la bonne ville de Lyon à la fonction *Contexte*. Les 19 passages où le corpus de Rabelais en fait mention sont restitués en 4 secondes (figure 13). Il en faut 13 pour quérir un mot plus fréquent, qui est familier à Rabelais sans être inconnu à Lyon, le mot *vin* (200 occurrences, 95 000 caractères). Pour une concordance, quelle que soit la fréquence, il suffit d'une ou deux secondes.

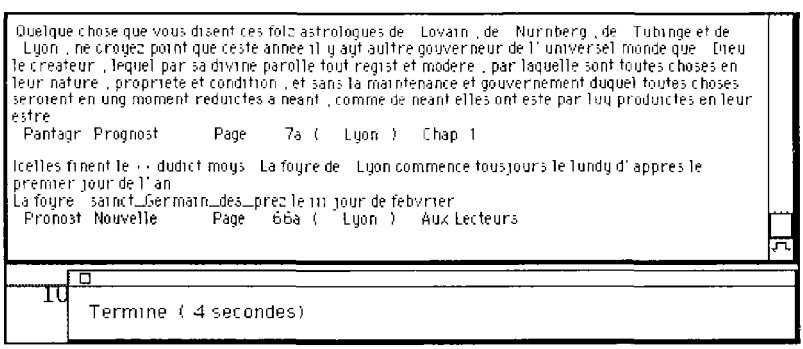

FIGURE 13 : La fonction *Contexte* dans le CD-ROM Rabelais
(Extrait des 19 emplois du mot Lyon).

L'originalité du produit ne tient pas seulement aux méthodes techniques, que l'utilisateur a le droit d'ignorer, mais aux fonctions disponibles dont certaines sont nouvelles. Outre les outils documentaires et statistiques habituels, le CD-ROM Rabelais offre une comparaison sous forme synoptique de plusieurs éditions du même texte, du moins là où les variantes sont les plus intéressantes, c'est-à-dire dans le cas du *Pantagruel* et du *Quart Livre*. Un clic sur n'importe quel mot de l'une des versions renvoie au mot correspondant de l'autre version. Voir figure 14.

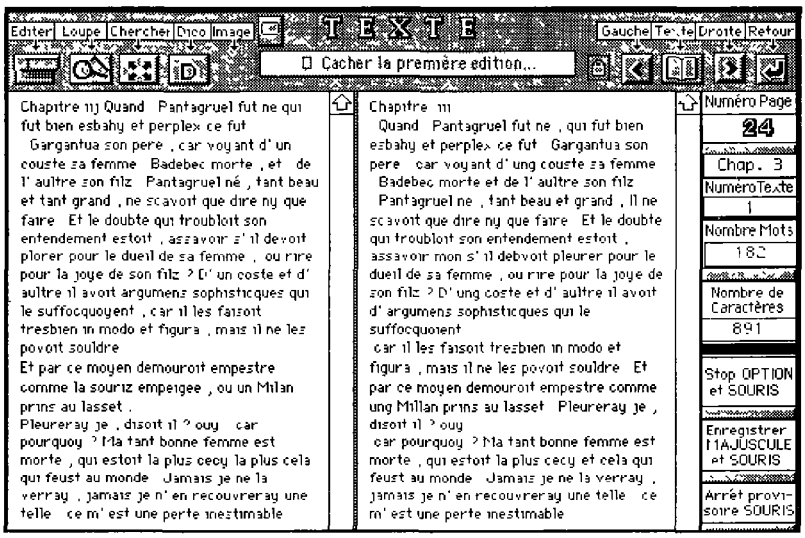

FIGURE 14 Comparaison de deux éditions du *Pantagruel*

En outre certains mots du texte de Rabelais ont été mis en relation avec les dictionnaires ou glossaires de l'époque. Ils apparaissent en caractères gras et réagissent au clic de la souris, montrant la définition du *Thresor* de Jean Nicot ou le commentaire de la *Brève déclaration*. L'italique désigne par ailleurs d'autres mots ou expressions qui bénéficient d'une explication et d'une illustration. Et de la même façon le clic sur le passage en italique fait apparaître successivement l'une et l'autre. Quatre cents documents iconographiques qui ont un rapport avec le texte du *Gargantua* ont été fournis par la Bibliothèque municipale de Lyon. Ils datent tous de l'époque de Rabelais et éclairent certains aspects méconnus du texte. Ces illustrations accompagnent les pages du *Gargantua* dans leur défilement mais elles constituent aussi une base de données autonome qu'on peut consulter librement. L'exemple de la figure 15 est emprunté à un très bel ouvrage de botanique qui avait cours au temps de Rabelais et que présente la figure 16.

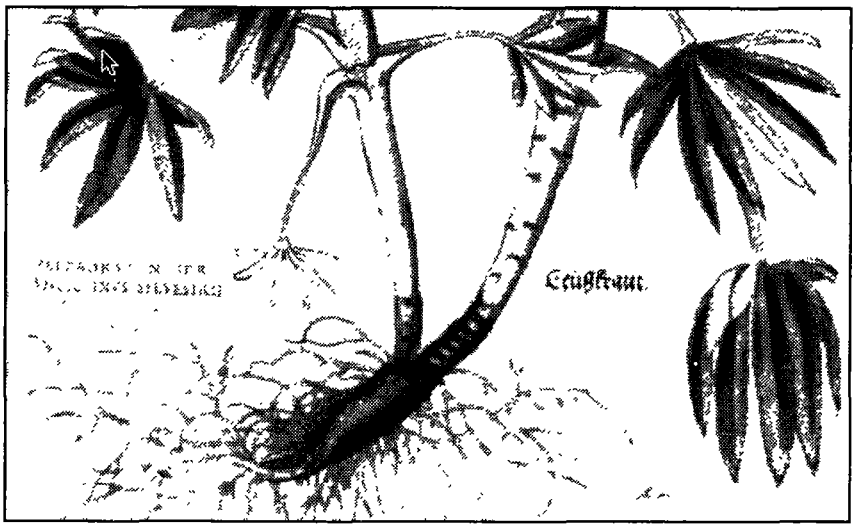

FIGURE 15 : Un document iconographique (extrait).

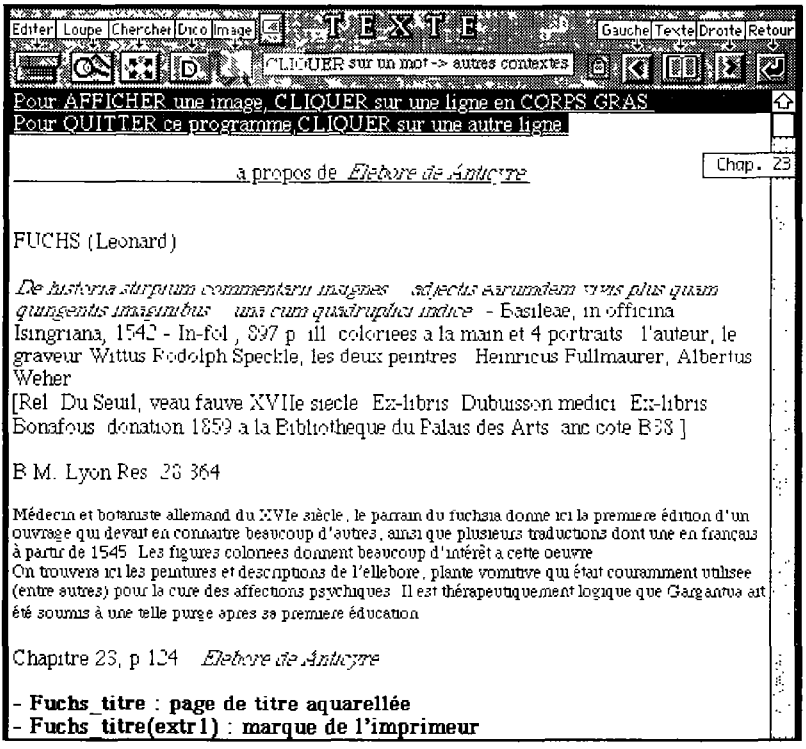

FIGURE 16 : Le commentaire de l'image.

Les facilités multimédia du CD-ROM ont été mises à profit, non seulement pour le traitement de la couleur (pourquoi s'en priver sur écran puisque le noir et blanc n'est pas plus économique), mais aussi pour l'incrustation de séquences animées et sonores qui expliquent et illustrent le fonctionnement de la base. Quand un bouton fait mystère, l'utilisateur a le moyen d'en exiger l'explication, brève ou détaillée. Un écran lui est d'abord montré avec un commentaire approprié et si cela ne suffit pas une séquence *Quicktime* lui est proposée, qui décompose les phases de l'opération, selon l'invitation reproduite dans la figure 17.

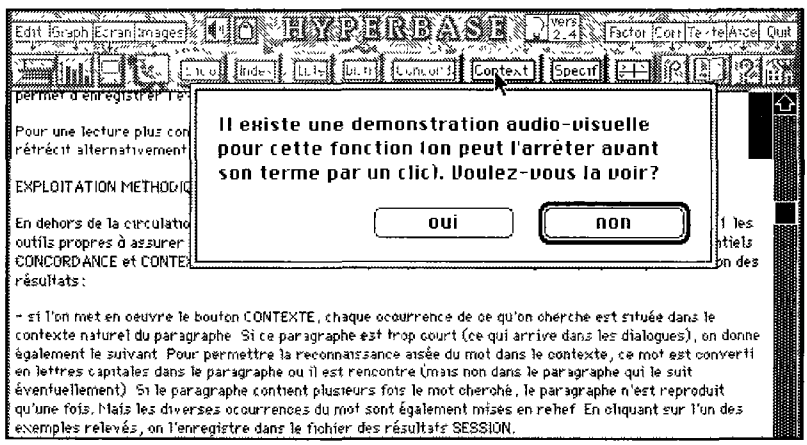

FIGURE 17 : Les aides multimédia (la fonction *Contexte* expliquée).

4. La version Internet d'*Hyperbase*

La version *Internet* d'*Hyperbase* pallie certaines des lacunes du CD-ROM. Même si le *CD-ROM Rabelais* est bi-standard, il n'offre pas les mêmes fonctionnalités sous *Windows* que sur matériel *Apple*. Et si les fichiers dérivés de la base (dictionnaires, index et concordances) ou associés à son emploi (documents iconographiques et leurs commentaires) sont accessibles sur *PC*, l'usage interactif de la base elle-même est réservé aux possesseurs de machines *Apple*. Quand au contraire on interroge une base sur le réseau, les résultats sont strictement les mêmes quel que soit le poste d'interrogation, qui peut être aussi bien une station *Unix*, un compatible *Dos* ou *Windows* et bien entendu une machine *Apple*. Pour atteindre ce haut degré de compatibilité, il faut passer par un langage précis et contraignant qui impose un codage particulier des messages à transmettre, dont chaque élément est doté de balises univoques. Ce langage *HTML*, instanciation de la norme plus générale *SGML*, gouverne les communications asymétriques entre le serveur et le client. Le client envoie des ordres brefs : clic sur un bouton, sur l'item d'un menu déroulant ou sur une zone soulignée (cela s'appelle une ancre ou un lien) et dans certains cas écrit un mot ou une expression dans une zone précise du dialogue. Sur un ordre exprès (bouton *submit* ou *OK*) ce message est transmis au serveur sous forme de chaîne de caractères, avec une ponctuation spécifique qui permet d'isoler et d'interpréter les paramètres. La réponse du serveur peut être immédiate s'il s'agit de transmettre un fichier déjà existant, qui peut être une page d'information, une image ou un fichier composite. Elle peut être légèrement différée si un traitement préalable est nécessaire, comme c'est le cas avec notre base Rabelais. Une

concordance, une recherche de contexte, un graphique, une liste de mots ou une analyse factorielle sont nécessairement des travaux sur mesure qu'il faut exécuter sur commande, au lieu que les documents iconographiques, les spécificités et certains résultats relatifs à la structure lexicale sont catalogués dans des fichiers tout prêts à l'emploi et immédiatement transmissibles. Dans tous les cas, à l'aller comme au retour, les données transmises doivent respecter les conventions du langage et en particulier coder les caractères accentués selon une norme commune à tous les standards.

Il serait oiseux d'entrer plus profondément dans les détails techniques. Peu importe quelles transformations ont été apportées au logiciel d'interrogation. Les plus importantes sont dues à une plus grande lourdeur des échanges quand s'interpose la distance. En traitement local, le dialogue peut être aussi vif ou laconique qu'on le souhaite. On peut répondre par oui ou par non et progresser rapidement dans la suite des répliques. À distance les répliques se transforment en tirades et le rythme des échanges se ralentit. Il faut quelques secondes pour obtenir une réponse et, dans l'attente du résultat, l'utilisateur est plongé dans un trou noir où il n'a d'autre alternative que d'attendre ou de suspendre. Afin de réduire ce silence, nous avons volontairement limité à 60 secondes le temps d'un échange, ce qui interdit les demandes impudentes ou imprudentes dont le résultat se ferait attendre trop longtemps (par exemple la concordance de tous les mots qui finissent par la lettre *s*). De toutes façons on a fixé à 1 000 le nombre maximum de lignes dans une concordance et à 100 000 caractères la taille de tout fichier transmis. Ces bornes raisonnables ont été établies pour éviter qu'un client gêne les autres ou se gêne lui-même par maladresse. On trouvera ci-dessous (figure 18) l'écran qui accueille l'utilisateur quand il prend contact avec la base à l'adresse :

(http ://ancilla.unice.fr/rabelais.html) ou (http ://134.59.31.3/rabelais.html)

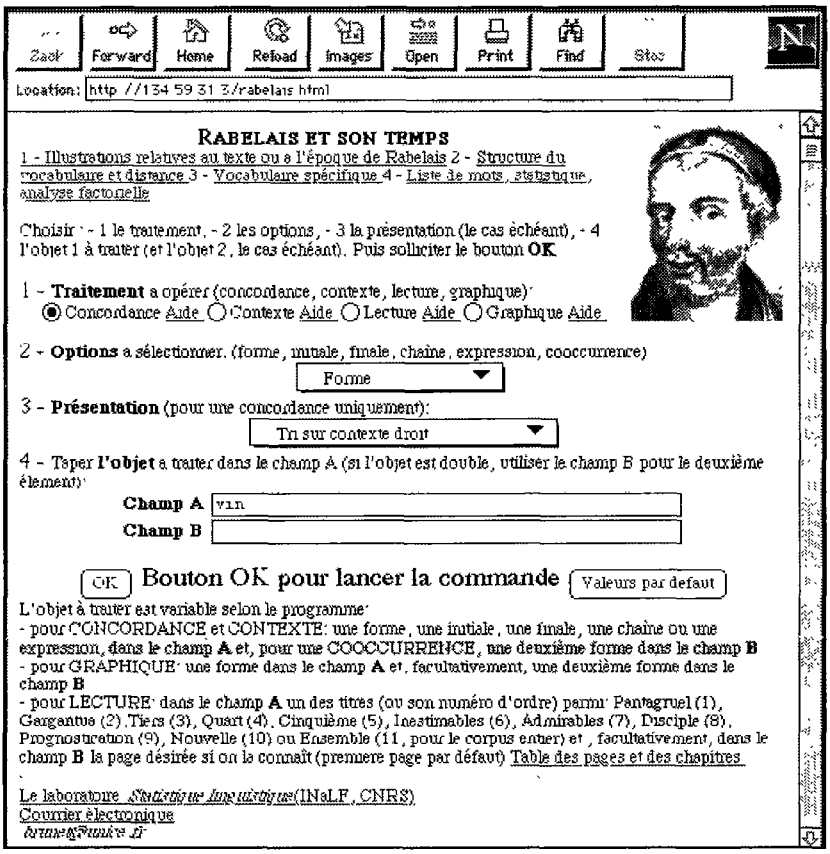

FIGURE 18 : La page d'accueil de la base *Rabelais* sur *Web*

Dans cet exemple, les paramètres sont réglés pour obtenir la concordance complète du mot *vin* dans le corpus avec une présentation ordonnée sur l'environnement à droite. Il ne faut guère que deux secondes d'exécution pour la collecte des 200 contextes de ce mot, à quoi s'ajoute le temps du transcodage et de la transmission, variable selon le type des liaisons.

```
CI 153a|                c' est à dire , en vin verité . Les deux parties estoient
PA 108a|      d' un grand hanat plein de vin vermeil , disant . Compere tout
DI  33f|    le monde . ? Le tiers est de vin vermeil qui passe en bonte tous les
CI 146a|signification evidente .  que le vin vos est en mespris , et par vos
CI  62d| . ¶Le cordieu vous aurez vostre vin à ceste heure :  je le vous promets
CI 188b|      estoit bouteille Pleine de vin a un aureille  ¶De vin , je dis ,
CI 180a|. C' est , dist frere  Jean , du vin à une aureille . Puis le vestit d'
CI 184a|   , et naturel flascon plein de vin  Phalerne ' lequel elle fist tout


Paramètres:
Concordance Forme vin( tridroit)
Temps d'exécution : 2 seconde(s)
Nombre d'appels de cette base: 912
```

FIGURE 19 : Extrait de la concordance triée du mot *vin*.

Outre la fonction *Concordance*, la page d'accueil présente les fonctions *Contexte*, *Lecture* et *Graphique* qui sont en tous points semblables à celles du CD-ROM et qui s'appliquent pareillement non seulement aux formes isolées mais aussi à une expression et à tout paradigme fondé sur une initiale, une finale ou une chaîne quelconque.

La visualisation des documents iconographiques est aussi simple et presque aussi rapide que dans la version CD-ROM. Un index (figure 20) est proposé qui en donne la liste en rattachant chaque illustration à son commentaire et à la page du *Gargantua* qu'elle complète. Ainsi le document précédemment reproduit dans la figure 15 – il s'agit de l'ellébore noir – est accessible à la ligne 230 de l'index, et de la même façon la page 124 du *Gargantua* à laquelle est liée cette illustration (figure 21) et la description bibliographique de l'ouvrage de botanique qui traite de l'ellébore et qui fait l'objet de la figure 16.

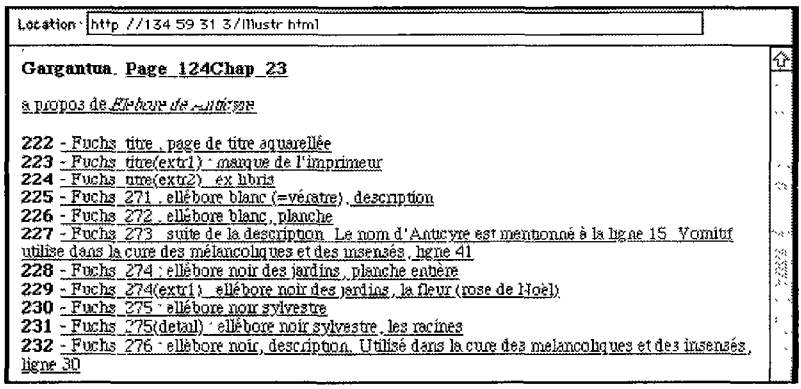

FIGURE 20 : L'index des illustrations.

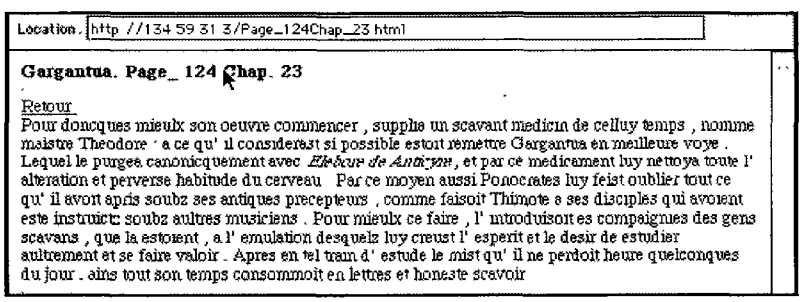

FIGURE 21 · La page 124 du *Gargantua* (chapitre 23).

Quant aux traitements statistiques, ceux qui ont trait au vocabulaire spécifique ou à la structure lexicale ont été préétablis et leur résultat peut être immédiatement transmis en sollicitant les ancres du haut de l'écran. Par contre, les traitements libres, soumis à la discrétion de l'utilisateur, font l'objet d'un écran particulier dont les options sont assez complexes. Elles permettent de constituer des tableaux à deux dimensions, avec les mots en ligne et les textes en colonne, et de les soumettre à l'analyse factorielle, avec différents types de pondération. La fonction *Graphique* a ici des possibilités étendues et s'applique aussi bien aux colonnes (profil d'un texte) qu'aux

lignes (distribution d'un mot). Elle s'applique également aux totaux marginaux comme aux individus. On prendra garde toutefois à ne pas allonger démesurément les listes, car la limite des 60 secondes veille là aussi à prévenir les excès.

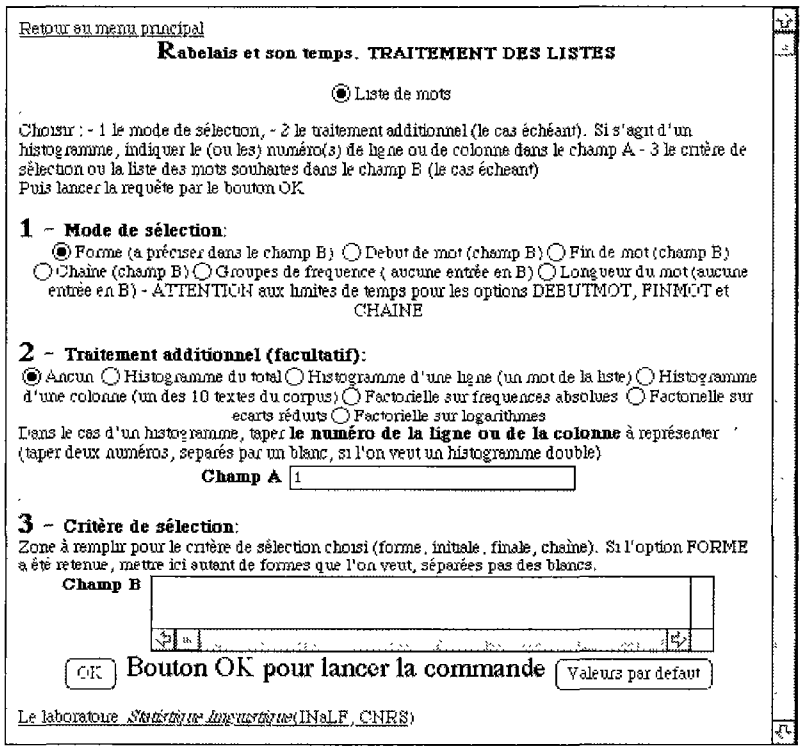

FIGURE 22 · Les traitements statistiques.

Conclusion

Il n'est pas certain que ces quatre variations d'un même thème aient épuisé les virtualités. Tout d'abord mille autres approches sont possibles que d'autres logiciels ont mises en œuvre avec succès. Mais aussi d'autres modes de diffusion peuvent être envisagés, qui appartiennent déjà aux opérations de routine. Rien n'est plus commun de nos jours que le téléchargement des fichiers et des logiciels. *Hyperbase* dans sa version à tout faire pourrait comme beaucoup d'autres être proposé en « shareware ». Mais, même sans donnée aucune, son poids reste dissuasif, avec deux millions d'octets à transmettre en format compressé. Et nos deux serveurs, qui ne disposent que d'un débit modeste de 65 kilobits, pourraient s'étouffer dans ce rude transfert. Une autre solution serait d'offrir un travail à façon, délocalisé, les données venant de l'extérieur sous forme de fichier *ASCII* à traiter, et les résultats retournant à l'envoyeur après traitement. Mais de tels services exigent des serveurs puissants et serviables et cela dépasse provisoirement les moyens du serviteur dévoué qui signe ces lignes.

Réseaux sémantiques et dictionnaires bilingues électroniques

Thierry FONTENELLE

Université de Liège, Belgique

1. Introduction

Dans cet article, je me propose d'aborder le problème de la réutilisation de ressources lexicales bilingues disponibles sur support informatique dans la perspective de la création et de l'amélioration de la composante lexicale de systèmes de traitement automatique du langage naturel. Le problème de la réutilisation des ressources lexicales existantes est en effet un problème épineux, tout particulièrement pour les ressources en langue française. Depuis le début des années 1980, les dictionnaires électroniques attirent l'attention des chercheurs en linguistique computationnelle. Ces derniers voient en effet dans les dictionnaires commerciaux électroniques une source inespérée de données lexicales d'ordre morphologique, syntaxique et sémantique cruciales pour des applications aussi diverses que la traduction automatique, la recherche documentaire, les interfaces de bases de données en langage naturel ou encore l'enseignement assisté par ordinateur. Depuis une dizaine d'années, les congrès et autres ateliers se sont multipliés et ont rassemblé des chercheurs du monde entier avides de partager leurs expériences, leurs succès, leurs frustrations aussi, dans l'extraction et la formalisation des informations lexicales contenues dans ces dictionnaires électroniques (voir entre autres, Boguraev & Briscoe, 1989 ; Zernik, 1991 ; Byrd, 1989 ; Atkins & Zampolli, 1994 ; Zampolli *et al.*, 1994). Force est de constater, cependant, que la grande majorité des efforts fournis dans le domaine concernent directement l'anglais. Les raisons en sont multiples. Il est intéressant de les analyser brièvement.

Il est incontestable que la tradition britannique des dictionnaires d'apprenants a joué un rôle considérable dans ce contexte. La richesse des dictionnaires monolingues pour apprenants étrangers comme le *Longman Dictionary of Contemporary English* (LDOCE, Procter, 1978), le *Collins Cobuild English Language Dictionary* (Sinclair, 1987), l'*Oxford Advanced Learner's Dictionary of English* (OALD, Cowie, 1989) ou, tout récemment, le *Cambridge International Dictionary of English* (CIDE, Procter, 1995) ne pouvait laisser indifférents tous ceux qui cherchent à automatiser la création,

la gestion et la mise à jour des lexiques informatisés utilisés dans des applications toujours plus sophistiquées. Les concepteurs sont en effet face à un cruel dilemme : soit ils font appel à des équipes de lexicographes spécialisés pour élaborer les lexiques dont ils ont besoin, ce qui, eu égard à la taille de ces lexiques pour des applications opérationnelles, s'avère extrêmement coûteux, soit ils essaient d'automatiser, ou du moins de semi-automatiser, le processus d'acquisition lexicale en examinant les possibilités d'extraction de ces informations de dictionnaires existants[1]. Or, il se fait que tous les dictionnaires mentionnés ci-dessus se distinguent des dictionnaires dits de langue par le recours systématique à un codage extrêmement élaboré des informations morphologiques, syntaxiques et parfois même sémantiques. Ainsi, des codes grammaticaux permettent de rendre compte de façon très détaillée de l'environnement syntaxique des items lexicaux, indiquant par exemple que tel verbe se construit avec une proposition infinitive en anglais, qu'il admet également les subordonnées introduites par *that* ou que tel nom, dans son sens indénombrable, gouverne la préposition *on*. Dans le cas de certains dictionnaires, on note même la présence de descriptions sémantiques s'apparentant aux règles de sélection. Ainsi, le LDOCE possède, dans sa version informatisée, une hiérarchie de codes sémantiques permettant de spécifier le trait sémantique inhérent d'un nom (par exemple, [+humain]) ou les contraintes sémantiques imposées par un verbe sur ses arguments (par exemple, le verbe X requiert un objet direct [+abstrait]).

La tradition lexicographique britannique n'a malheureusement pas eu d'émules dans le monde francophone. On ne connaît à ce jour aucun dictionnaire français disponible dans le commerce et possédant une richesse grammaticale comparable. Certains projets se sont penchés sur l'exploitation du dictionnaires Zyzomys de Hachette (Bouchard & Emirkanian, 1994 ; Bouchard *et al.*, 1991 ; Ide *et al.*, 1994), un dictionnaire de langue disponible sur CD-ROM. Il faut néanmoins constater que les quelques fragments de taxonomies ainsi que les descriptions morphosyntaxiques extraits de ce dictionnaire souffrent difficilement la comparaison avec les résultats obtenus sur les dictionnaires anglais. Si l'on considère que certains mettent en doute l'utilité même des recherches effectuées ces 15 dernières années sur les dictionnaires anglais (Veronis & Ide, 1994), on peut raisonnablement affirmer que les tentatives de réutiliser les dictionnaires monolingues français n'ont pas révolutionné la lexicographie computationnelle, principalement par manque de ressources adéquates.

Les considérations qui précèdent ont toutes trait aux tentatives de réutilisation de dictionnaires monolingues. Les dictionnaires bilingues, quant à eux, ont été encore plus négligés. Une des premières raisons est que les bandes magnétiques des dictionnaires bilingues comportant le français comme langue cible ou source n'ont été mises à la disposition que de quelques centres de recherche. Le *Robert & Collins dictionnaire anglais-français, français-anglais* (Atkins & Duval, 1978), par exemple,

1 Je passe sous silence le rôle non négligeable joué par les corpus de textes dans la problématique de l'acquisition lexicale. Il est évident que les dictionnaires informatisés ne peuvent fournir qu'une fraction de l'information lexicale nécessaire à un système élaboré du TALN. Les chercheurs se sont donc tout naturellement également tournés vers les ressources textuelles dont le traitement statistique permet de dégager des descriptions lexicales reflétant mieux l'usage, la fréquence d'emploi, la distribution, etc. À priori, la situation du français devrait être meilleure puisque la création de corpus ne dépend pas d'une quelconque tradition lexicographique anglo-saxonne, comme c'est le cas pour les dictionnaires. Dans les faits, cependant, on remarque que la plupart des logiciels permettant de traiter « en profondeur » un corpus (analyseurs syntaxiques) ne sont valables que pour les corpus anglais

n'est disponible, dans sa version non abrégée, que dans notre propre laboratoire à Liège ainsi qu'au Lexical Systems Group d'IBM à Yorktown Heights (Byrd, 1989 ; Byrd *et al.*, 1987 ; Boguraev, 1991). D'autres groupes ont eu accès à des versions de poche des dictionnaires bilingues publiés par Collins (le groupe de l'ISSCO par exemple, cf. Petitpierre *et al.*, 1994 et Robert, 1995 pour la version anglais-français ou le groupe de Pise pour l'anglais-italien – cf Picchi *et al.*, 1992), mais la petite taille de ces dictionnaires ne les rend pas susceptibles d'un traitement intéressant dans le cadre des recherches qui nous occupent ici, à savoir le dépistage et la formalisation des collocations.

Les dictionnaires informatisés bilingues ont été quelque peu négligés pour une autre raison, plus fondamentale. En règle générale, en effet, le format de ces dictionnaires est beaucoup moins structuré que celui des dictionnaires monolingues anglais dont il est question plus haut. Alors que ces derniers se rapprochent des bases de données lexicales dont les linguistes ont besoin pour leurs travaux (leur organisation logique permettant plus facilement l'identification de chaque information – partie du discours, définition, exemple, codes grammaticaux...), les dictionnaires bilingues comme le Robert & Collins ne sont le plus souvent que des dictionnaires lisibles par machine, en ce sens que les fichiers qui sont mis à la disposition des chercheurs ne sont rien d'autre que les bandes magnétiques ayant servi à la photocomposition de l'ouvrage. La structure logique des fichiers est dès lors limitée à la spécification de codes permettant le changement de typographie (passage à l'italique, etc.). La transformation de ces fichiers en véritables bases de données est loin d'être chose aisée et réclame une analyse approfondie de la microstructure des entrées. La nouvelle génération de dictionnaires bilingues, basée sur la norme SGML, devrait faciliter l'exploitation de ces ouvrages de référence en permettant l'identification immédiate des éléments composant les entrées lexicales (à ce sujet, on lira avec intérêt les résultats obtenus dans le cadre du projet COMPASS coordonné par le Centre de Recherche de Rank Xerox à Grenoble : le but de ce projet est d'exploiter les versions informatisées de dictionnaires bilingues pour construire la composante lexicale d'un système interactif d'aide à la compréhension de textes ; outre le dictionnaire Collins-Klett anglais-allemand dont il est question ailleurs dans ce volume – cf. Segond & Breidt – le projet COMPASS utilise la version informatisée du récent dictionnaire Oxford-Hachette (Corréard & Grundy, 1994), le tout premier dictionnaire bilingue construit à partir d'un corpus et balisé à l'aide du langage SGML – cf. Bauer *et al.*, 1995 ; Segond & Zaenen, 1994).

2. Construction d'une base de données lexico-sémantique à partir du dictionnaire Robert & Collins

Les travaux présentés dans cet article ont été réalisés dans le cadre d'un doctorat en linguistique anglaise (Fontenelle, 1995). L'idée de base était de réutiliser la version lisible par machine de la partie anglais-français du dictionnaire Robert & Collins afin d'exploiter l'information collocationnelle qu'il contient. Ces recherches ont été entreprises suite à la constatation que l'appareil métalinguistique du Robert & Collins s'avère être une véritable mine de renseignements sur les propriétés combinatoires des items lexicaux. Le traitement explicite des restrictions quant aux sujets et objets des verbes, par exemple, rend ce dictionnaire bilingue extrêmement utile comme source

d'information pour la construction semi-automatique d'une base de données bilingue de collocations.

L'appareil métalinguistique utilisé par les lexicographes du Robert & Collins couvre toute une série d'informations cruciales pour la désambiguïsation. Ces indications, données en italiques, vont de la spécification de la partie du discours (*n, adj, vt, vi...*) aux codes matières (*Bio, Comput, St Ex* [Stock Exchange], *Mus...*), en passant par les niveaux de langues (*fml, infml, fig, lit*) et les restrictions de sélections et autres collocations. C'est à cette dernière catégorie que je me suis plus particulièrement intéressé, dans le but de rendre ces contraintes collocationnelles accessibles à l'utilisateur humain ou à la machine. Le système appliqué par les lexicographes mérite d'être présenté et illustré afin de donner par la suite une idée plus claire des potentialités de la base de données.

- Les noms sujets typiques d'un verbe apparaissent entre crochets.
- Les noms typiquement utilisés comme compléments d'un autre nom apparaissent également entre crochets.
- Les objets directs d'un verbe et les noms typiquement modifiés par un adjectif apparaissent à côté de la partie du discours (pas de crochets ni de parenthèses).
- Les adjectifs, verbes et adverbes modifiés par un adverbe apparaissent sans crochets ou parenthèses.

Les exemples suivants illustrent cette pratique :

Objets typiques
arouse *vt* **(b)** (*cause*) *suspicion, curiosity etc* éveiller, susciter; *anger* exciter, provoquer; *contempt* susciter, provoquer.
dissipate *vt fog, clouds, fears, suspicions* dissiper; *hopes* anéantir; *energy, efforts* disperser, gaspiller; *fortune* dissiper, dilapider.
harbour 3 *vt* **(b)** *suspicions* entretenir, nourrir; *fear, hope* entretenir.

Sujets typiques
billow 2 *vi [sail]* se gonfler; *[cloth]* onduler; *[smoke]* s'élever en tourbillons *or* en volutes, tournoyer.
flap 3 *vi* **(a)** *[wings]* battre; *[shutters]* battre, claquer; *[sails]* claquer.
puff up 1 *vi [sails etc]* se gonfler; *[eye, face]* enfler.

Combinaisons N+N
flap 1 *n* **(a)** *[wings]* battement, coup; *[sails]* claquement...
beard *n [fish, oyster]* barbe; *[goat]* barbiche; *[grain]* barbe, arête...

Combinaisons Adj+N
baseless *adj accusation etc* sans fondement; *suspicion* sans fondement, injustifié.
well-founded *suspicion* bien fondé, légitime.
well-grounded *suspicion* bien fondé, légitime.

Les entrées ci-dessus appellent plusieurs commentaires. Tout d'abord, il est clair que l'accès à l'information dépend du public visé par le dictionnaire. Dans le cas présent, nous avons affaire à un dictionnaire qui permet à un utilisateur francophone de déterminer le sens exact, et par conséquent, la traduction d'un mot anglais en fonction

de son contexte. Ce contexte est décrit par le lexicographe à l'aide du vocabulaire métalinguistique en italiques dont la richesse et l'abondance contribuent à la qualité de l'ouvrage. Il s'agit donc, dans la perspective de l'utilisateur francophone, d'un dictionnaire de décodage. Le même utilisateur qui souhaiterait se servir de cette partie du dictionnaire pour encoder, c'est-à-dire générer, du texte en anglais se verrait confronté au problème crucial de l'accès aux collocations. Si l'on prend comme hypothèse de base que, comme le souligne Hausmann (1979), les collocations sont des combinaisons polaires comportant une base et un collocatif, la première étant responsable de la sélection du second dont le sens dépend de sa mise en contexte avec la base, on s'aperçoit aisément que, dans les entrées ci-dessus, la base est l'élément en italiques alors que l'entrée proprement dite constitue le collocatif. L'espace me manque ici pour reprendre tous les arguments en faveur de la création de dictionnaires combinatoires où les informations collocationnelles seraient classées sous la base, et non, comme c'est le cas ici, sous le collocatif (voir à ce sujet Hausmann, 1985, 1989 ; Cowie, 1986 ; Heid, 1994). Les informations présentées ici ne permettent pas de répondre facilement à des questions lexicographiquement aussi intéressantes que :

- Que peut-on faire à un 'soupçon' ? (≈ Quels verbes peuvent prendre le mot *suspicion* comme objet direct ?)
- Quelles peuvent être les caractéristiques d'un 'soupçon' ? (≈ Quels adjectifs peuvent qualifier le mot *suspicion* ?)
- Que peut faire une voile ? (≈ Quels verbes peuvent prendre le mot *sail* comme sujet ?)

Ces questions sont en fait celles auxquelles le lexicographe est confronté dans son travail quotidien. On notera également que ces problèmes sont au cœur de nombreuses recherches actuellement menées en linguistique informatique où l'acquisition de collocations par des méthodes statistiques dans des corpus de textes est un sujet brûlant (on lira avec intérêt les travaux de Smadja, 1991, 1993 ; Grefenstette, 1994a et b ; Church & Hanks, 1990 ; Church *et al.*, 1994 ; Zernik, 1991). Les applications en sont aussi bien la construction de nouvelles ressources lexicographiques que l'élaboration de lexiques pour la génération automatique du langage naturel (Smadja, 1993).

Dans le cas qui nous occupe, la recherche des collocations se fait non pas dans des corpus de textes, mais dans la version informatisée du dictionnaire Robert & Collins qui est lui-même considéré comme un corpus pré-digéré d'informations lexicales. Comme on l'a vu plus haut, les collocations pertinentes que nous recherchons sont présentes dans le dictionnaire, mais l'organisation même de celui-ci, par le classement alphabétique des collocatifs, ne permet pas à l'utilisateur de découvrir le lien étroit qui unit des entrées comme *dissipate, arouse, harbour, well-founded* ou *baseless*. La présence de l'item *suspicion* dans la micro-structure de chacune de ces entrées permet cependant à une machine de reconstruire l'environnement collocationnel d'une base donnée en extrayant les occurrences de cette base et les entrées sous lesquelles elle apparaît. Pour autant que le dictionnaire soit organisé en base de données, il est alors possible de repérer toutes les occurrences du mot *suspicion* en italiques dans la totalité du dictionnaire. Un utilisateur qui n'aurait que la version papier du dictionnaire à sa disposition devrait alors parcourir les 800 pages de la partie anglais-français, alors que cette requête ne prend que quelques fractions de seconde dans notre base de données. La liste résultant de cette interrogation pour le nom *suspicion* comprend les entrées suivantes :

arouse, avert, awake, baseless, confirm, dissipate, drive away, eliminate, entertain, harbour, just, quieten, remove, rest, rouse, suspicious (2x), *suspiciously* (2x), *suspiciousness* (2x), *unsuspicious* (2x), *verify, well-founded, well-grounded.*

Il ne m'est pas possible, pour des raisons d'espace, de décrire les problèmes liés à la transformation de la bande magnétique du Robert & Collins en une base de données permettant des accès multiples à l'information lexicale. Il importe néanmoins de noter que cette transformation fut loin d'être triviale et qu'elle a occupé Jacques Jansen, informaticien dans notre département, pendant de nombreux mois. Le modèle relationnel a été choisi pour diverses raisons, entre autres parce que d'autres dictionnaires, par exemple le LDOCE (Procter, 1978), étaient déjà disponibles dans le même format, ce qui facilite les fusions et applications se servant de ressources multiples. Le Robert & Collins est actuellement disponible sur PC et sur station UNIX (Sun Sparc Station) et des programmes d'application, écrit en C par Luc Alexandre et en Clipper par moi-même, permettent une interrogation souple de la base de données. Certaines critiques ont été émises à l'encontre du modèle relationnel (Boguraev *et al.*, 1992), mais, comme la base de données comporte toute une série de tables distinctes (pour les entrées, les parties du discours, les traductions, les indicateurs métalinguistiques, les exemples, etc.), le fait qu'une entrée comporte trois traductions alors qu'une autre n'en comporte qu'une n'est en fait pas un problème. Comme les différentes tables sont liées par des champs communs, la redondance est évitée et les programmes d'application sont chargés de reconstruire, à l'intention de l'utilisateur, une vue non fragmentaire des données lexicales. Des arguments plus détaillés en faveur de l'utilisation du modèle relationnel sont proposés par Michiels (1995). La structure proprement dite de la base de données du Robert & Collins est, quant à elle, décrite en détail dans Fontenelle (1995) et dans certains rapports internes du projet DECIDE dans lequel s'inscrit ce travail (Jansen & Fontenelle, 1994).

3. Enrichissement de la base de données

Dès le départ, il est apparu qu'il serait souhaitable de pouvoir structurer les informations collocationnelles de notre base de données. En effet, si l'utilité des renseignements repris plus haut sur les possibilités combinatoires du mot *suspicion* est incontestable, il n'en reste pas moins que les relations lexico-sémantiques unissant ce mot aux entrées sous lesquelles on le trouve sont de nature hétérogène. Ainsi, les verbes *harbour* (entretenir, nourrir) et *dissipate* (dissiper) peuvent tous deux prendre *suspicion* comme objet direct mais, du point de vue sémantique, renvoient à des sens diamétralement opposés. De la même façon, *baseless* (sans fondement) et *well-founded/ well-grounded* (légitime) sont, dans le contexte de *suspicion*, liés par une relation d'antonymie. Comme le principe même d'une base de données est de permettre à l'utilisateur non seulement d'extraire de l'information par des chemins d'accès divers, mais aussi de mettre à jour les données et de les enrichir, il fut décidé d'ajouter systématiquement, pour l'ensemble du vocabulaire métalinguistique (collocations et restrictions de sélection), un code permettant de formaliser la relation lexico-sémantique existant entre la base et le collocatif. Pour ce faire, je me suis basé sur la Théorie Sens⇔Texte d'Igor Mel'čuk et plus particulièrement sur le mécanisme des fonctions lexicales. Ce mécanisme est, je crois, suffisamment connu pour m'épargner une nouvelle introduction au domaine (on consultera les trois volumes du *Dictionnaire explicatif et combinatoire du français contemporain* – Mel'čuk *et al.*, 1984, 1988, 1992

– ainsi que les contributions d'Apresjan, Tutin, Sérasset et Alonso-Ramos & Mantha dans ce volume). Je me contenterai de rappeler ici que le terme 'fonction lexicale' désigne une relation de sens assez abstraite telle que l'expression linguistique de cette fonction dépend du lexème auquel elle vient se joindre. La notation traditionnelle, empruntée au modèle mathématique, est $f(x) = y$, où f est la fonction lexicale (FL), x est le mot-clé et y la valeur de la fonction. L'exemple typique est la fonction Magn, qui dénote l'intensification ou le degré élevé de quelque chose, comme dans Magn (*célibataire*) = *endurci* ou Magn (*menteur*) = *fieffé*. On a également des fonctions censées codifier les relations de verbes supports, c'est-à-dire les verbes pratiquement vides de sens qui se combinent avec le mot-clé, comme $Oper_1$ (*attention*) = *faire*, $Oper_1$ (*crime*) = *commettre* ou $Oper_1$ (*question*) = *poser*. On notera que les fonctions lexicales présentent l'ensemble de la cooccurrence lexicale restreinte d'un lexème donné. Dans le cas des collocations, le mot-clé est la base au sens où l'entend Hausmann (1979) et la valeur de la fonction est le collocatif. Le modèle de Mel'čuk, qui comprend une soixantaine de fonctions lexicales de base, permet en outre de formaliser des relations sémantiques d'ordre paradigmatique, et non pas seulement syntagmatique (par exemple, des relations de synonymie (FL=Syn), d'antonymie (FL=Anti), de dérivation morphologique (A_0(*loi*)=*légal*), etc.).

La majeure partie de la tâche à laquelle je me suis attelé pendant environ deux ans a été de déterminer la nature de la relation lexico-sémantique unissant les 70 000 indicateurs métalinguistiques en italiques et l'entrée sous laquelle on trouve ces indicateurs. Pour ce faire, j'ai écrit un programme d'application me permettant de créer une nouvelle table pour accueillir cette relation sémantique. Le processus d'encodage, guidé par une série de menus déroulants destinés à réduire les risques d'incohérence, a permis de créer, pour chaque item lexical apparaissant dans le vocabulaire métalinguistique du dictionnaire, une sorte de réseau sémantique où la base est liée aux autres mots par une relation empruntée au modèle mel'čukien. Dans le cas du mot *suspicion* illustré plus haut, cela signifie que j'ai attribué, pour chacune des occurrences de ce mot dans une entrée du dictionnaire (*arouse, awake, baseless, harbour*, etc.), une étiquette correspondant à une fonction lexicale. De plus, comme le dictionnaire est un ouvrage bilingue, j'ai également traduit l'item en italiques afin de créer une véritable base de données lexico-sémantique permettant d'accéder à l'information collocationnelle par le biais de l'anglais ou du français (la base – l'indicateur à l'origine en italiques, sa traduction française, ajoutée manuellement, le collocatif – l'entrée anglaise ou la traduction française donnée par le dictionnaire). Le réseau sémantique pour les 26 termes gravitant autour de *suspicion* se présente comme suit, en respectant la notation proposée par Mel'čuk, à savoir f(x) = y :

$causfunc_0$ (suspicion / soupçon) = arouse (éveiller) [S]
liqu (suspicion/soupçon) = avert (écarter) [S]
$causfunc_0$ (suspicion/soupçon) = awake (éveiller) [S]
antiver (suspicion/soupçon) = baseless (sans fondement) [S]
$real_1$ (suspicion/soupçon) = confirm (confirmer) [S]
liqu (suspicion/soupçon) = dissipate (dissiper) [S]
liqu (suspicion/soupçon) = drive away (chasser) [S]
liqu (suspicion/soupçon) = eliminate (éliminer) [S]
$oper_1$ (suspicion/soupçon) = entertain (nourrir) [S]
$oper_1$ (suspicion/soupçon) = harbour (entretenir) [S]
ver (suspicion/soupçon) = just (justifié) [S]
liqu (suspicion/soupçon) = quieten (calmer) [S]
liqu (suspicion/soupçon) = remove (dissiper) [S]

(suspicion/soupçon) = rest (fonder) [S]
$causfunc_0$ (suspicion/soupçon) = rouse (éveiller) [S]
a_1 (suspicion/soupçon) = suspicious (soupçonneux) [P]
a_2 (suspicion/soupçon) = suspicious (suspect) [P]
adv_1 (suspicion/soupçon) = suspiciously (avec méfiance) [P]
adv_2 (suspicion/soupçon) = suspiciously (d'une manière suspecte) [P]
s_0qual_1 (suspicion/soupçon) = suspiciousness (caractère soupçonneux) [P]
s_0qual_2 (suspicion/soupçon) = suspiciousness (caractère suspect) [P]
$antia_1$ (suspicion/soupçon) = unsuspicious (peu soupçonneux) [P]
$antia_2$ (suspicion/soupçon) = unsuspicious (qui n'a rien de suspect) [P]
$real_1$ (suspicion/soupçon) = verify (vérifier) [S]
ver (suspicion/soupçon) = well-founded (bien fondé) [S]
ver (suspicion/soupçon) = well-grounded (bien fondé) [S]

Sans entrer dans les détails, on remarque que ce réseau sémantique contient des informations collocationnelles cruciales pour l'encodage. Ainsi, la fonction **Liqu** (<'liquider', 'détruire'>) permet de répondre à la question : Quels verbes expriment la 'suppression' d'un soupçon ? (*avert, dissipate, drive away, eliminate, quieten, remove* pour l'anglais ; *écarter, dissiper, chasser, éliminer, calmer* pour le français). La fonction **$Oper_1$**, qui correspond à la notion de verbe 'support' pratiquement vide de sens, ou du moins sémantiquement appauvri, se retrouve dans les combinaisons *entertain_suspicion* ou *harbour_suspicion* (on *entretient* ou on *nourrit* des soupçons à l'égard de quelqu'un ≈ on soupçonne quelqu'un). La fonction **Ver**, qui permet de générer des adjectifs signifiant que le mot-clé est 'correct' ou 'tel qu'il doit être', permet de codifier le lien unissant *suspicion* et *just, well-founded* ou *well-grounded.* Inversement, **Ver** se combine avec la FL **Anti** pour former la fonction complexe **Anti-Ver** et ainsi étiqueter la relation unissant *suspicion* et *baseless.*

L'élément apparaissant entre crochets à la fin de chaque ligne correspond au codage de la nature typographique de l'élément en italiques. Ainsi, [S] signifie que le mot suspicion apparaît à ce que nous avons appelé le niveau de 'surface' (d'où le 's') dans les entrées *arouse, avert, awake,* c'est-à-dire sans parenthèses ni crochets. Comme je l'ai indiqué plus haut, cette information typographique est cruciale puisqu'elle permet de déterminer la nature du lien syntaxique entre l'indicateur métalinguistique et son entrée. Ici, le niveau de surface [S] correspond au statut d'objet direct pour un verbe transitif (something *arouses* somebody's suspicions – quelque chose *éveille* les soupçons de quelqu'un). Ce même niveau [S] code le lien entre un adjectif et le nom qu'il modifie (*just* suspicions – des soupçons *justifiés*). Dans le cas présent, nous n'avons pas de verbes intransitifs dont *suspicion* pourrait être le sujet. Si tel était le cas, l'indicateur apparaissant alors entre crochets, l'information typographique serait représentée par [C]. La formalisation de cette information dans la base de données permet de poser des questions comme, par exemple, « Quels sont les verbes transitifs qui peuvent avoir l'item X comme sujet/objet possible ? ».

On remarque en outre que, dans certains cas, l'information typographique indique que le terme *suspicion* apparaît au niveau [P], c'est-à-dire entre parenthèses dans la version imprimée du dictionnaire. Les parenthèses sont en fait utilisées pour ajouter des micro-définitions. On a par exemple l'entrée suivante :

unsuspicious *adj* *(feeling no suspicion)* peu soupçonneux, peu méfiant
(causing no suspicion) qui n'a rien de suspect, qui n'éveille aucun soupçon

Si l'on considère que le mot *suspicion* implique la présence de deux actants (la personne qui soupçonne et ce qui est soupçonné), on arrive au cas de figure suivant où A_1 et A_2 sont les fonctions lexicales utilisées par Mel'čuk pour générer l'adjectif correspondant au premier et au second actant respectivement :

$Anti A_1$ (suspicion) = unsuspicious (peu soupçonneux)
$Anti A_2$ (suspicion) = unsuspicious (qui n'a rien de suspect)

Il ne s'agit bien sûr plus ici d'information collocationnelle puisque la relation est de nature paradigmatique. Comme elle est présente dans le dictionnaire, cependant, il aurait été dommage de s'en passer et, dans le cadre d'un réseau sémantique et d'une application possible dans la recherche documentaire, ce genre de relation se doit d'être formalisé. On notera également qu'il a été possible de semi-automatiser l'attribution de ces fonctions lexicales grâce à une analyse des structures récurrentes dans les micro-définitions. Ainsi, les structures « causing + N » ou « feeling + N » sont utilisées comme des indices permettant de déterminer la nature de la relation lexico-sémantique en question. Je me suis inspiré ici des travaux d'Amsler (1980), Michiels & Noël (1984) ou Ahlswede & Evens (1988) pour identifier les formules définitoires récurrentes dans les micro-définitions du Robert & Collins.

4. Collocations, terminologie et fonctions lexicales

Le Robert & Collins n'est pas un dictionnaire permettant d'étudier les langues de spécialité. Il ne contient donc normalement que des informations sur la langue générale, même si le vocabulaire plus technique est loin d'en être absent. La base de données que nous avons construite permet cependant de transformer le dictionnaire en une sorte de thésaurus où l'encyclopédique se mêle au lexical. Comme le note Frawley (1988), les lexiques spécialisés ne fournissent généralement pas d'information sur l'environnement collocationnel des termes (même si cette situation est en train de changer ; cf. Cohen, 1986 ; Verlinde *et al.*, 1992). C'est pourtant de cette information que les traducteurs, rédacteurs et apprenants ont besoin. Pour formaliser le discours spécialisé utilisé pour parler d'un terme donné, les théories de Mel'čuk ne sont probablement pas les plus appropriées parce qu'elles ne permettent de coder que des relations standard de la langue générale et les langues de spécialité ont le plus souvent recours à des relations très spécifiques (cf. Blampain *et al.*, 1992 et Blampain, 1993 qui identifient un certain nombre de relations notionnelles spécifiques à un domaine donné et pour lesquelles le modèle de Mel'čuk n'offre aucun équivalent). On constate néanmoins que la base de données collocationnelle du Robert & Collins contient des informations précieuses sur la combinatoire de certains termes et, d'un point de vue phraséologique, permet de répondre à certaines questions fréquemment posées par les terminologues travaillant sur des corpus spécialisés. Si l'on considère que notre base de données, grâce aux nombreux index sur tous les types d'information, permet d'accéder aux données via n'importe quelle clé, séparément ou en combinaison, il est possible de formuler des questions combinant syntaxe et sémantique dont les réponses nécessiteraient une conversation approfondie avec un spécialiste du domaine, comme le souligne Frawley (1988).

Prenons l'exemple du mot anglais *sail* (fr. voile) qui, nous l'avons vu plus haut, peut se combiner avec des verbes tels que *flap, billow, puff up* et bien d'autres encore.

L'analyse des occurrences de *sail* dans les rubriques métalinguistiques des entrées du Robert & Collins nous permet d'obtenir une image assez détaillée du discours nautique utilisé pour parler des voiles en anglais et en français. La formalisation des collocations à l'aide des fonctions lexicales permet de poser les questions suivantes à notre base de données :

Q : Y a-t-il un terme désignant un ensemble régulier de voiles ?
A : Mult (sail) = set (Fr. jeu)

Q : Y a-t-il un verbe désignant le son typique des voiles ?
A : Son (sail) = flap (Fr. claquer)

Q : Y a-t-il un nom désignant le son typique des voiles ?
A : S_0Son (sail)= flap (Fr. claquement)

Q : Y a-t-il des verbes transitifs prenant le mot *sail* comme objet direct et signifiant que le but typique associé aux voiles est réalisé ?
A : $Real_1$ (sail) = get up, haul up, hoist, put up, shake up, spread (Fr. hisser, déployer)

Q : Y a-t-il des verbes transitifs prenant le mot *sail* comme objet direct et signifiant que le but typique associé aux voiles n'est plus réalisé ?
A : $AntiReal_1$ (sail) = haul down, (Fr affaler), lower (Fr. abaisser), strike (Fr. amener)

Q : Quelles sont les parties typiques d'une voile ?
A : Part (sail) = belly (Fr. creux), gore (Fr. pointe)

Q : La voile fait-elle partie d'une entité plus importante ?
A : Whole (sail) = barge (Fr. barge), yacht (Fr. yacht, voilier)

On notera que les deux dernières relations ne sont pas couvertes par le modèle mel'čukien parce qu'elles ne correspondent pas à des fonctions à proprement parler, mais plutôt à des relations de 1 → n. Dans la base de données, j'ai introduit les relations **Part** et **Whole** pour rendre compte de liens méronomiques présents dans le dictionnaire. Même si ces relations sont de nature sémantique, et non lexicale, elles n'en sont pas moins cruciales pour la recherche documentaire ou pour l'analyse automatique du langage[2].

2. L'analyse des constituants d'une phrase et la levée d'ambiguïtés potentielles peuvent être grandement facilitées si les entrées lexicales comportent une indication des relations « a_comme_partie » (correspondant à notre **Part**), « fait_partie_de » (**Whole** dans notre base de données) ou « a_comme_instrument_typique » (correspondant à la fonction lexicale $\mathbf{S_{instr}}$ chez Mel'čuk) Considérons les deux phrases suivantes
(a) *John painted the jeep with the new wheels.* (John a peint la jeep aux roues neuves)
(b) *John painted the jeep with the new brush* (John a peint la jeep avec le nouveau pinceau)
Il est clair que [the jeep with the new wheels] ne forme qu'un seul et même constituant alors qu'on a deux constituants distincts, [the jeep] et [with the new brush] (attaché au prédicat *paint*), dans (b) Ce type d'analyse ne peut reposer que sur l'indication explicite des liens suivants dans le lexique (Gener est également une fonction lexicale qui renvoie à l'hyperonyme du mot-cle)
S_{instr} (paint) = brush ..
Part (vehicle) = wheel...
Gener (jeep) = vehicle

Il faut admettre que l'enrichissement de la base de données à l'aide des fonctions lexicales a ses limites. Je n'ai pas essayé à tout prix d'imposer une fonction lexicale pour modéliser une relation trop spécifique. Plutôt que d'attribuer une FL qui ne correspondrait que de loin à la relation unissant une base et son collocatif, j'ai préféré m'abstenir dans un certain nombre de cas. J'ai également évité de multiplier les FL pour ne pas créer des relations qui ne seraient utilisées que deux ou trois fois dans une base de données contenant, rappelons-le, plus de 70 000 enregistrements. L'absence de fonction lexicale standard pour certaines combinaisons n'en empêche pas moins de compléter le réseau sémantique évoqué ci-dessus. Ainsi, on peut poursuivre le dialogue imaginaire relatif au nom *sail* et poser les questions suivantes à la base de données :

Q : Y a-t-il d'autres verbes transitifs pouvant prendre *sail* comme objet direct et indiquant ce que l'on peut faire à une voile ?
A : *bend* (Fr. enverguer), *furl* (Fr. ferler), *gore* (Fr. mettre une pointe à), *puff* (Fr. gonfler), *reef* (Fr. prendre un ris dans), *swell* (Fr. gonfler), *trim* (Fr. gréer).

Q : Y a-t-il des verbes prenant *sail* comme sujet et indiquant ce qu'une voile peut faire ?
A : Les voiles peuvent *billow* ou *billow out* (Fr. se gonfler), *fill out* (Fr. gonfler), *puff out* ou *puff up, swell* ou *swell out* (Fr. se gonfler), elles peuvent aussi *gibe* (passer d'un bord à l'autre du mât).

Q : Y a-t-il des combinaisons Adj+N ou N+N servant à décrire les voiles ? (par ex. quels adjectifs peuvent qualifier le nom *sail* ?)
A : Les voiles peuvent être *billowy* (Fr. gonflé par le vent), *full* (Fr. plein) ou *swelling* (Fr. gonflé). Le syntagme « the *billow* of a sail » (Fr. gonflement) est également attesté, à côté d'autres collocations N+N correspondant à des fonctions lexicales standard comme *flap of sails* ci-dessus (S_0 Son).

5. Conclusions

Comme on a pu le constater, l'appareil métalinguistique du Robert & Collins offre un point de départ intéressant pour la construction de réseaux sémantiques et les combinaisons reprises plus haut attestent de la richesse de ce dictionnaire. Ces combinaisons représentent ce que l'on serait tenté d'appeler « du matériau linguistique d'aide à la décision » puisqu'elles permettent de rafraîchir la mémoire parfois défaillante d'un traducteur ou d'activer une série de termes parmi lesquels l'utilisateur du dictionnaire électronique pourra opérer un choix selon le sens qu'il souhaite rendre. Les différentes clés d'index de la base de données permettent d'interroger celle-ci via plusieurs chemins d'accès, qu'il s'agisse de la base anglaise de la collocation (l'item en italiques), de sa traduction (ajoutée manuellement), du collocatif (l'entrée anglaise du dictionnaire), du collocatif français ou de la fonction lexicale. Comme les clés d'accès peuvent être combinées, il est possible de poser des questions sémantiquement très complexes (par exemple, quels sont les verbes supports de type inchoatif pouvant prendre comme objet direct le nom anglais *habit* ? ⇔ IncepOper_1 (habit) = *acquire, contract, develop, form, take to* – fr. *prendre, contracter).* On se rapproche dès lors du dictionnaire des collocations tel que l'envisage Hausmann (1979, 1985)

puisque l'utilisateur module à son gré des requêtes portant sur un réseau lexico-sémantique où le nœud est la base de la collocation et les flèches unissant ce nœud aux collocatifs et termes associés correspondent aux fonctions lexicales mel'čukiennes (enrichies d'autres relations non prévues par le modèle standard).

La souplesse des programmes d'application et l'ajout de l'information lexico-sémantique a radicalement modifié la conception du dictionnaire bilingue. Alors que l'utilisateur était au départ soumis aux contraintes inhérentes de l'ordre alphabétique, il peut à présent utiliser la base de données comme un thésaurus informatisé couplé à un dictionnaire combinatoire. Le linguiste dispose alors d'un outil souple lui permettant de tester, sur un vaste échantillon des vocabulaires anglais et français, des hypothèses relatives à la structure des lexiques de ces deux langues. Le traducteur et l'apprenant disposent quant à eux d'une ressource lexicale bilingue leur permettant de poser des questions qu'ils n'avaient peut-être pas encore imaginées et dont les réponses devraient faciliter l'encodage. Dans une perspective purement computationnelle, enfin, la base de données lexicale du Robert & Collins devrait pouvoir être utilisée pour désambiguïser les textes et donc, dans une certaine mesure, résoudre l'épineux problème de la détermination du sens exact d'un mot dans son contexte (et de sa traduction si l'on se place dans un contexte multilingue)[3].

Remerciements

Les recherches décrites dans cet article s'inscrivent dans le cadre d'une thèse de doctorat soutenue par l'auteur en octobre 1995. Elles s'inscrivent également dans le cadre des activités du projet de recherche DECIDE (*Designing and Evaluating Extraction Tools for Collocations in Dictionaries and Corpora* – MLAP 93/19) financé en partie par la Commission des Communautés européennes[4]. Elles n'auraient pas abouti sans la contribution essentielle de Jacques Jansen qui a analysé avec minutie les bandes magnétiques du dictionnaire Robert & Collins et conçu la base de données relationnelle. Qu'il en soit vivement remercié, ainsi que Luc Alexandre, qui a écrit certains des programmes d'application pour PC et pour stations UNIX. Il faut en outre souligner que ce travail n'aurait pu être réalisé sans la confiance des éditeurs (Harper Collins Publishers et les Dictionnaires Le Robert) qui ont accepté de mettre à notre disposition la version électronique de leurs dictionnaires à des fins de recherche fondamentale. Ils méritent également notre gratitude.

3 Pour rester dans le domaine nautique que nous avons évoqué plus haut, songeons simplement à la polysémie du verbe *amener* dans « Le matelot amène la voile » et « Le matelot amène ses enfants à la voile »

4 Outre le département d'anglais de l'Université de Liège (coordinateur du projet), le consortium du projet DECIDE comprend l'Institut fur maschinelle Sprachverarbeitung de l'Université de Stuttgart et le Centre de Recherche Rank Xerox de Grenoble B T S Atkins (anciennement éditeur en chef du Robert & Collins) participe également au projet en tant que consultante

Une maquette de base lexicale multilingue à pivot lexical (« acceptions multilingues ») : PARAX

Étienne BLANC

GETA-CLIPS, Institut IMAG (UJF & CNRS), Grenoble, France

• *Abstract* •

An Acception Based Multilingual Lexical Database has been designed under Hypercard. The aim of this mock-up, which contains about 200 words in five languages (English, French, German, Russian and Chinese), is to contribute to the validation of the concept of Multilingual Acceptions in the building of Multilingual Lexical Databases.

1. Introduction

On sait l'importance majeure qu'ont pris les dictionnaires dans les systèmes de Traduction Automatique ou les systèmes d'Aide à la Traduction, et le développement consécutif récent des travaux dans le domaine des ressources lexicales réutilisables et notamment des bases lexicales multilingues.

Le GETA qui est engagé dans un projet multilingue de Traduction Automatique Fondée sur le Dialogue, le projet LIDIA (Boitet & Blanchon, 1993 ; Blanchon, 1994a et 1994b), a naturellement été amené lui aussi à s'intéresser aux bases lexicales multilingues (Sérasset, 1994a et 1994b).

Notre réflexion a surtout porté sur un type particulier de structure de base multilingue, la structure de type pivot lexical ou pivot « par acceptions ». Malgré l'intérêt de ce type de structure, il n'a été encore que peu utilisé, une exception notable étant le système de TAO « ULTRA » de l'Université du Nouveau Mexique (Farwell, Guthrie & Wilks, 1993).

Nous présentons, dans cet article, une maquette de base lexicale de structure pivot « par acceptions » vue dans la perspective de l'utilisateur. Le but de cette maquette étant de permettre une expérimentation de ce type de structure du point de vue

linguistique, nous nous sommes efforcés de la concevoir simple d'emploi et facilement extensible à un grand nombre de langues (elle en comporte cinq actuellement : allemand, anglais, chinois, français et russe). Pour le moment, nous n'avons pas attaché trop d'importance ni à la structure linguistique, relativement rudimentaire, ni à la couverture lexicale (environ 500 acceptions actuellement).

Le choix d'Hypercard pour la réalisation est, entre autres, motivé par la grande souplesse de réalisation que permet ce logiciel. Nous pouvons ainsi faire évoluer notre base pour l'adapter à des représentations linguistiques variées, ce qui nous a en particulier permis d'intégrer des descriptions lexicales et sémantiques utilisées dans Mel'čuk (1988) dont l'extension multilingue nous paraît particulièrement intéressante.

2. La structure pivot « par acceptions »

2.1. Bases multilingues de type transfert et de type pivot

On peut schématiquement envisager deux types de structure pour une base multilingue traitant n langues :

- ou bien n dictionnaires monolingues et $n(n-1)$ liaisons unidirectionnelles (ou $n(n-1)/2$ bidirectionnelles) entre ces dictionnaires deux à deux ; c'est la structure multibilingue, encore appelée « de transfert » par analogie avec la technique de TAO de même nom ;

- ou bien n dictionnaires monolingues et un dictionnaire d'un langage pivot (ou interlingue), et $2n$ liaisons unidirectionnelles (ou n bidirectionnelles) entre ces dictionnaires et le dictionnaire pivot ; c'est la structure dite « pivot ».

La seconde approche semble la plus appropriée à des applications typiquement multilingues, ne serait-ce qu'en raison du nombre moindre de liaisons interdictionnaires à réaliser et c'est effectivement celle qui a été utilisée pour des projets comme ULTRA (Farwell, Guthrie & Wilks, 1993) et CICC (Yaoliang & Zhendong, 1991) qui font intervenir un grand nombre de langues (anglais, allemand, chinois, japonais et espagnol pour ULTRA, japonais, chinois, malais, indonésien et thai pour CICC).

Mais si la structure pivot semble la mieux adaptée à des applications multilingues, il reste qu'elle est de réalisation et de maintenance plus complexe que la structure multibilingue.

2.2. Interlangue ontologique et interlangue lexicale : les acceptions multilingues

Une des principales difficultés de l'approche pivot réside dans la réalisation de l'interlingue.

L'interlingue la plus séduisante, mais la plus complexe à réaliser, est l'interlingue ontologique (dictionnaire de concepts). C'est le choix adopté pour le plus grand projet mondial de base lexicale bilingue, le projet EDR (EDR, 1993), qui comporte un dictionnaire de 640 000 concepts. Notons cependant que, pour tenir compte

des difficultés d'une approche purement ontologique, EDR a adopté une structure mixte pivot-transfert : les dictionnaires monolingues anglais et japonais sont chacun mis en correspondance avec le dictionnaire de concepts (structure pivot), mais sont également mis en correspondance entre eux (structure de transfert).

À la complexité de l'interlingue ontologique s'oppose la simplicité de l'interlingue lexicale « par acceptions ». Dans cette approche, on ne cherche, en effet, pas à définir intrinsèquement les éléments de l'interlingue. Ces éléments, que nous nommerons ici « acceptions multilingues », ou par abréviation « acceptions », peuvent être définis de la façon suivante :

Un lemme donné d'une langue L a en général plusieurs sens, ou « acceptions monolingues ». Si l'on peut faire correspondre à une acception monolingue de ce lemme des acceptions monolingues sémantiquement identiques de lemmes dans les langues L', L'', etc. on dira que l'ensemble de ces acceptions monolingues représente l'acception multilingue associée à ces diverses acceptions monolingues. Le lexique « par acceptions » est constitué de l'ensemble de ces acceptions multilingues.

Une acception multilingue n'est donc pas définie intrinsèquement, comme l'est un concept dans une base ontologique. On peut cependant lui attribuer l'ensemble des propriétés sémantiques partagées par les acceptions monolingues qu'elle représente, on retrouve ainsi au moins en partie l'intérêt d'une base ontologique, tout en évitant la difficulté d'une description complète de concepts.

Cette première définition élémentaire de l'acception multilingue suppose que l'on puisse associer à une acception donnée dans une langue donnée des équivalents sémantiques dans toutes les langues de la base que l'on veut construire. Il est clair que la situation n'est pas toujours aussi simple, et que, si pour certaines acceptions monolingues (les termes techniques par exemple) la définition d'une acception multilingue ne présente pas de difficultés, pour d'autres une correspondance directe ne sera pas possible et on aura à considérer des correspondances plus complexes (analogues, par exemple, aux relations de sous-relation, super-relation et synonymie utilisées dans EDR). Et il n'est pas évident qu'une solution satisfaisante puisse toujours être trouvée.

C'est pourquoi il nous a paru important de tester expérimentalement la possibilité de construction d'une base de ce type sur un nombre suffisamment grand de langues appartenant à des familles différentes. C'est la raison d'être de la maquette que nous décrivons ici.

3. Structure générale de PARAX

Le dictionnaire pivot et les dictionnaires monolingues de PARAX sont chacun constitués d'une pile HYPERCARD (figure 1).

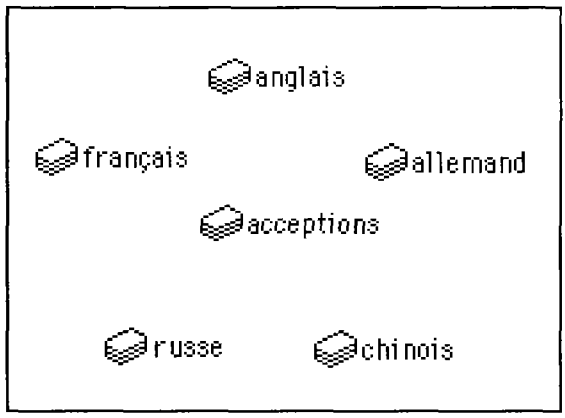

FIGURE 1 Les dictionnaires de PARAX.

Les dictionnaires monolingues sont actuellement au nombre de cinq (français, anglais, allemand, chinois et russe) et contiennent environ 200 mots chacun. Le dictionnaire pivot (dictionnaire d'acceptions) contient environ 500 acceptions.

Notons que l'ajout de nouvelles langues est chose aisée. Il suffit, pour créer un nouveau dictionnaire, de dupliquer une pile « patron » et de compléter un fichier descripteur. Les liens avec le dictionnaire d'acceptions sont créés automatiquement lors de l'entrée des termes.

Nous allons commencer la description de PARAX en y simulant une navigation, et pour cela, cliquons sur l'icône d'un dictionnaire monolingue, le dictionnaire français, par exemple.

4. Les dictionnaires monolingues de PARAX

On accède ainsi à la première page de ce dictionnaire (figure 2) qui présente la liste des lemmes indexés. En cliquant sur un élément de cette liste, le lemme *argent* par exemple, on accède à la page courante correspondante du dictionnaire (figure 3).

On voit que deux sens ont été retenus pour le lemme *argent*. On voit également que chacun de ces sens est décrit de façon autonome, morphosyntaxiquement dans la colonne de gauche, sémantiquement dans la colonne centrale. La séparation des descriptions morphosyntaxique et sémantique reflète le fait que cette dernière description, commune à tous les équivalents dans les autres langues, n'est qu'une copie provenant de la pile pivot.

D'autres informations peuvent apparaître sur demande : en cliquant sur le mot « FLEXICALES » qui suit la description morphosyntaxique d'un sens, on obtient la liste des fonctions lexicales de Mel'čuk. En cliquant sur le mot « EXEMPLES », on obtient des exemples d'utilisation du sens correspondant du lemme. Ces informations étant également accessibles à partir du dictionnaire d'acceptions, nous en parlerons plus en détail dans le paragraphe consacré à ce dictionnaire.

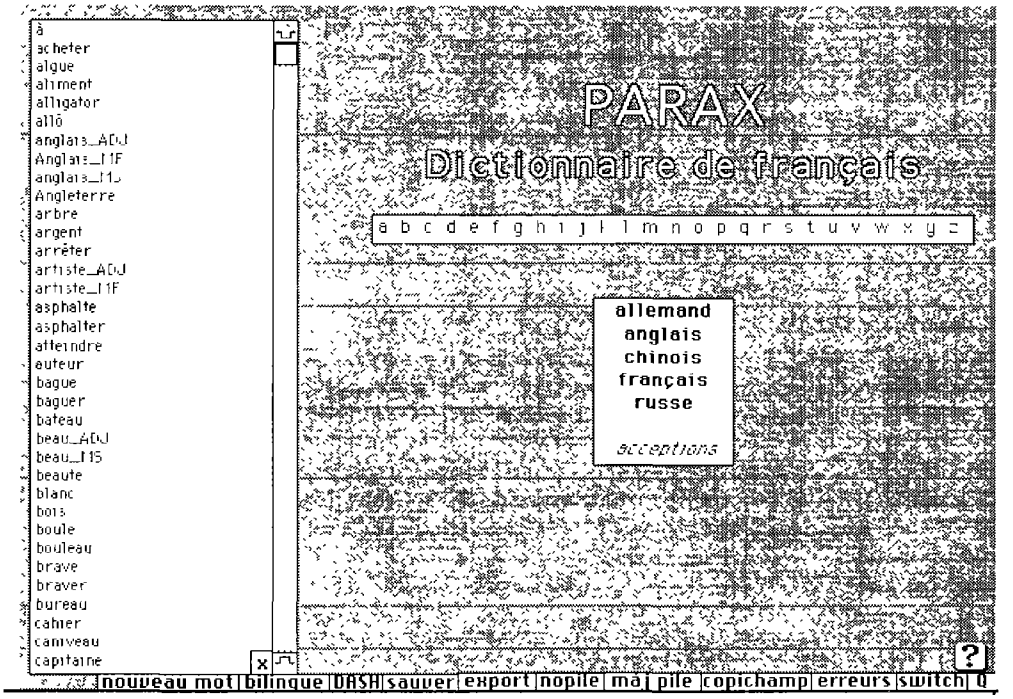

FIGURE 2 : La première page du dictionnaire français.

Dict français	argent	
argent #money	**#money**	
SENS 2 CAT nc GNR mas	{toute sorte de monnaie}	
[1IONOPIVOT/FLEXICALES/EXEMPLES]	SEMENT argent	
argent #silver_metal	**#silver_metal**	
SENS 1 CAT nc GNR mas	{metal noble de numéro atomique 47}	
[1IONOPIVOT/FLEXICALES/EXEMPLES]	SEMENT matiere	

FIGURE 3 · Le lemme *argent* dans le dictionnaire français.

Une carte donnée est consacrée à un lemme et non à une famille dérivationnelle ou unité lexicale. Cependant, les dérivations sont indiquées et le passage du lemme dérivé au lemme origine, ou inversement du lemme origine aux lemmes dérivés possibles, est immédiat (le passage a plus précisément lieu non pas entre lemmes origines et lemmes dérivés, mais entre acceptions de lemmes origines et acceptions de lemmes dérivés).

La figure 4a représente ainsi un fragment de la carte du lemme « beauté » décrivant l'une des trois acceptions retenues pour ce lemme : la beauté en tant que qualité esthétique. Le fait que cette acception (notée #beauté_esthet) dérive du lemme « beau » dans l'acception #beau_esthet, suivant le schéma adjectif → nom abstrait, est indiqué par la présence de la variable morphologique « DRVA » (dérivation de l'adjectif), par la valeur « nabst » (nom abstrait) de cette variable et par l'indication du lemme origine et de son acception dans la notation *£beau @beau_esthet*. Cette nota-

tion reflète le fait qu'en cliquant sur *@beau_esthet* on accède à la description de l'acception origine #beau_esthet sur la carte du lemme origine « beau » (figure 4b). Les mots composés et les locutions sont traités de façon analogue.

Dict. français	**beauté**
beauté #beaute_esthet	**#beaute_esthet**
SENS:1 CAT: vbimp DRVA: nabst CAT:	{qualité de ce qui fait éprouver une émotion
nc GNR fem. DRVA: nabst *£beau*	esthétique }
@beau_esthet	SEMENT: état.
[MONOPIVOT/FLEXICALES/EXEMPLES]	

FIGURE 4a : Dérivation adjectif → nom abstrait : en cliquant sur « @beau_esthet », on accède au lemme origine de la figure 4b

Dict. français	**beau**
beau #beau_esthet	**#beau_esthet**
SENS:1. CAT: adj. DRVAP: nabst. *£beauté*	{qui fait éprouver une émotion esthétique }
@beauté_esthet	SEMPROPR: qualif.
[MONOPIVOT/FLEXICALES/EXEMPLES]	

FIGURE 4b : Dérivation potentielle adjectif → nom abstrait.

Revenons maintenant à la carte relative au lemme « argent » de la figure 3 et cliquons sur le mot « MONOPIVOT » qui termine l'article relatif au sens 2, auquel correspond le code d'acception.multilingue #silver_metal. On accède alors à la carte du dictionnaire d'acceptions décrivant cette acception multilingue (figure 5).

5. Le dictionnaire d'acceptions

La figure 5 montre la carte du dictionnaire d'acceptions décrivant l'acception #silver_metal.

Source Français	FR	**#silver_metal**	Target
argent #silver_metal		**#silver_metal °AL °AN °CH °FR °RU**	
SENS 1 CAT nc GNR mas		{métal noble de numéro atomique 47}	
[MONOPIVOT/FLEXICALES/EXEMPLES]		SEMENT matière	

FIGURE 5 : Une carte du dictionnaire d'acceptions.

La colonne de gauche comporte la recopie de la description morphosyntaxique du représentant de cette acception dans le dictionnaire « source », c'est-à-dire le dictionnaire monolingue à partir duquel on vient d'accéder à la pile d'acceptions.

L'acception elle-même est décrite dans le champ central avec indication des langues dans lesquelles elle est représentée à l'aide des symboles AL (allemand), AN (anglais), CH (chinois), FR (français) et RU (russe). En cliquant sur l'un de ces symboles, on fait apparaître dans le champ de droite la description morphosyntaxique du représentant de l'acception dans la langue correspondante (figure 6).

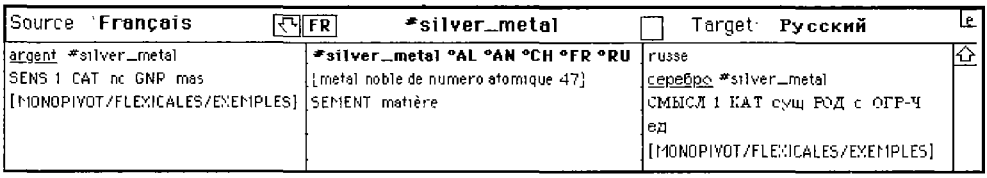

FIGURE 6 . La carte du dictionnaire pivot après choix du russe comme langue cible.

La figure 7 correspond au cas où la langue source est le chinois, les langues cibles sont multiples, et où l'on a fait apparaître les exemples dans les différentes langues, accessibles, comme nous l'avons dit, à partir des piles monolingues comme à partir de la pile d'acceptions. On peut de même faire apparaître les fonctions lexicales de Mel'čuk dans la ou les langues où elles ont été entrées.

On voit également, sur cette figure, que la définition sémantique est maintenant rédigée en chinois. Un bouton dans le bandeau supérieur de la carte permet, en effet, de choisir la langue de définition sémantique parmi toutes les langues de la base. Cette possibilité n'est pas uniquement une commodité d'utilisation pour des locuteurs de différentes langues, mais a un intérêt plus fondamental. En effet, si l'on est parvenu à donner d'une même acception des définitions correspondantes dans les diverses langues de la base et que ces définitions sont illustrées par des exemples identiques, ou au moins proches dans les diverses langues, on peut penser que l'on est parvenu à cerner correctement cette acception multilingue.

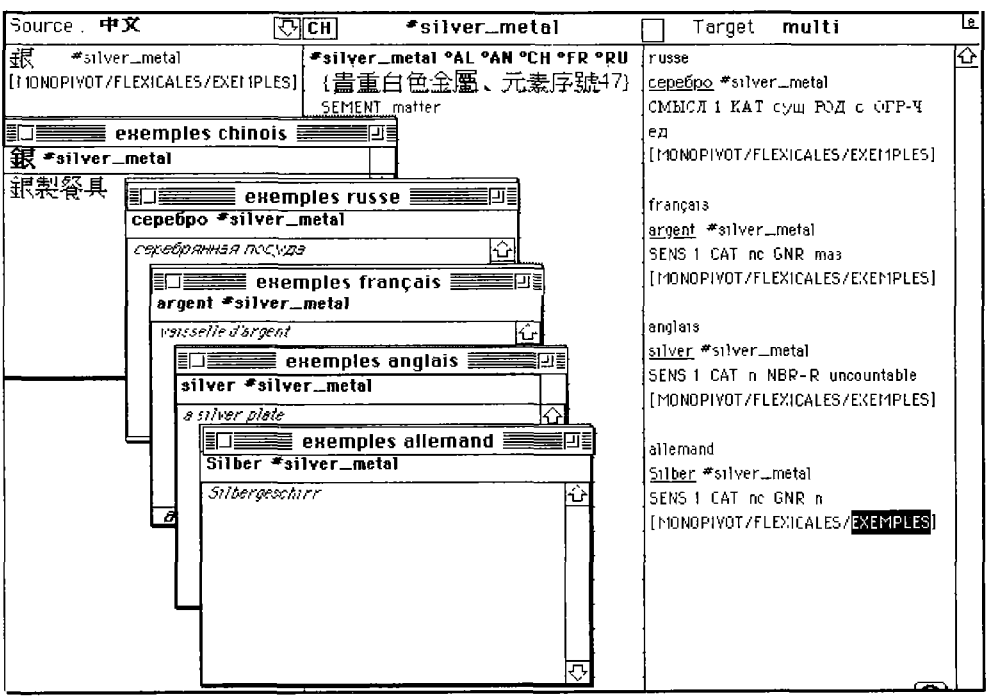

FIGURE 7 : Le chinois est choisi comme langue source et comme langue de description sémantique ; on a par ailleurs fait apparaître les exemples dans les différentes langues

Pour résumer cette présentation de la structure de la pile d'acceptions, on peut dire que cette pile stocke les informations sémantiques et propose, pour faciliter la consultation, des copies des informations lexicales dans les diverses langues (alors que les piles monolingues stockent les informations lexicales et proposent des copies des informations sémantiques).

5. Sur-acceptions et sous-acceptions

Il arrive bien entendu que l'on ait du mal à faire une correspondance suffisamment exacte entre des acceptions monolingues pour aboutir à une acception multilingue satisfaisante.

Un exemple élémentaire que nous avons traité correspond à la notion de sous-acception ou de sur-acception. Considérons, par exemple, le sens 2 du lemme français *capitaine* : « officier qui commande une compagnie d'infanterie, un escadron de cavalerie, une batterie d'artillerie » (figure 8a). Lorsqu'on passe au dictionnaire pivot (figure 8b), on voit que, si l'on choisit l'allemand comme langue cible, il existe un niveau de raffinement supérieur : avant de cliquer sur le symbole « AL », il faut choisir une des sous-acceptions de l'acception #captain_officer ; l'équivalent allemand viendra alors se placer dans le champ de droite en face de cette sous-acception.

Dict français	capitaine	
capitaine #captain_navy SENS 1 CAT nc GNP mas [MONOPIVOT/FLEXICALES/EXEMPLES]	**#captain_navy** {officier qui commande un navire de commerce ou de pêche}	
capitaine #captain_officer$ SENS 2 CAT nc GNP mas [MONOPIVOT/FLEXICALES/EXEMPLES]	**#captain_officer$** {officier qui commande une compagnie d'infanterie, un escadron de cavalerie, une batterie d'artillerie *artillerie/air/cavalerie/infanterie*}	
capitaine #captain_sportteam SENS 3 CAT nc GNP mas [MONOPIVOT/FLEXICALES/EXEMPLES]	**#captain_sportteam** {chef d'une équipe sportive}	

FIGURE 8a · Le lemme français *capitaine*.

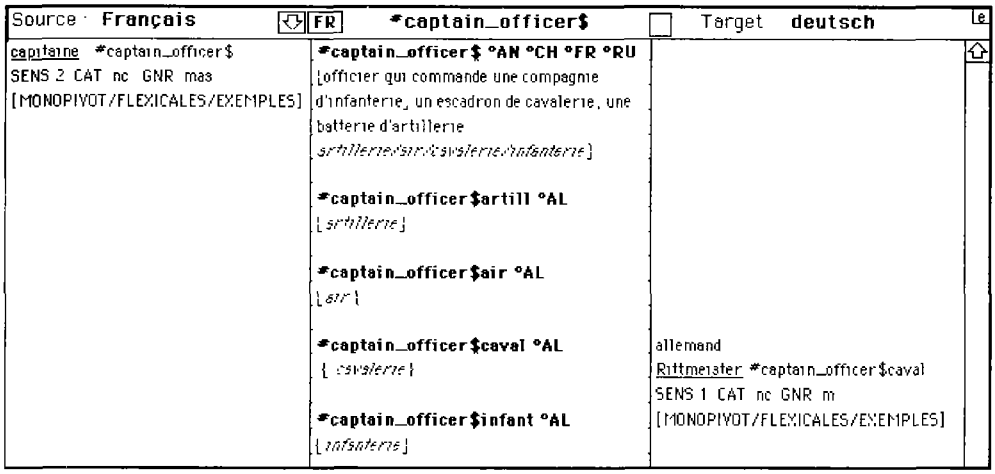

FIGURE 8b : Le passage d'une acception à une sous-acception.

6. Traitement des collocations

Considérons l'expression allemande *die Faust ballen*. Elle est décrite dans la carte *Faust* du dictionnaire monolingue allemand (figure 9a) avec une décomposition la rattachant à l'acception multilingue #formerboule du verbe *ballen*, (à laquelle, rappelons-le, on accède directement en cliquant sur *£ballen @formerboule* dans le champ de gauche).

L'expression anglaise correspondante *to clench the fist* est décrite dans la carte du dictionnaire monolingue anglais avec rattachement à l'acception multilingue #empoigner_serrer du verbe *to clench*.

Il est clair qu'il est préférable d'effectuer le passage de l'expression allemande à l'expression anglaise par l'intermédiaire d'une acception multilingue #serrer le poing (figure 9b), plutôt que de créer artificiellement une acception multilingue associant le verbe allemand *ballen* au verbe anglais *to clench*.

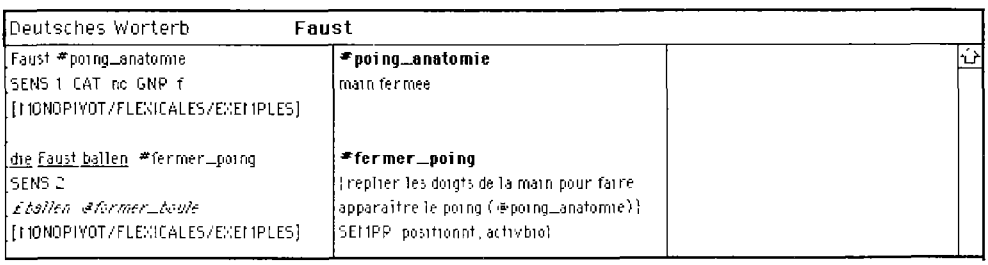

FIGURE 9a · L'expression idiomatique allemande *die Faust ballen*.

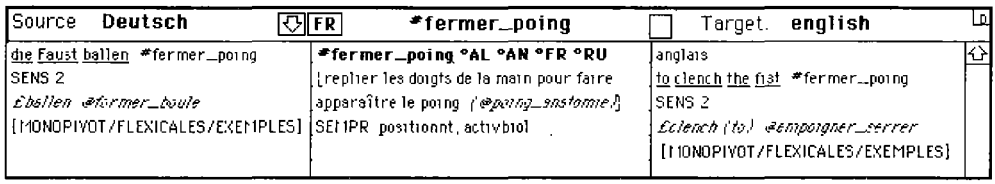

FIGURE 9b : Passage à l'expression idiomatique anglaise *to clench the fist.*

7. Exploitation de la base PARAX

Le principal objectif d'une base lexicale n'est en général pas de proposer une consultation directe de l'information qu'elle contient, mais de permettre d'utiliser cette information pour la réalisation d'outils lexicographiques, notamment de dictionnaires, dictionnaires destinés à une utilisation humaine ou à une application machine.

7.1. Outil d'aide à l'écriture de dictionnaires machine

Dans le cadre du projet de LIDIA de TAO multilingue fondée sur le dialogue, nous avons d'abord utilisé PARAX comme outil d'aide à l'écriture et à la maintenance des dictionnaires machine en l'associant à une interface hypertextuelle de documentation et de commande du générateur de systèmes de TAO ARIANE (l'interface DASH, pour Documentation Ariane sous Hypercard).

La figure 10 montre ainsi une copie d'écran comportant en arrière-plan une page du dictionnaire pivot de PARAX et, en premier plan, la page correspondante du fichier source d'un dictionnaire de transfert ARIANE tel qu'il apparaît dans l'interface DASH. Ce dictionnaire, qui a été envoyé à DASH par messagerie électronique depuis le serveur ARIANE, peut ensuite être modifié (ou même créé) en tenant compte des informations contenues dans PARAX, puis être réexpédié au serveur ARIANE pour compilation et exploitation.

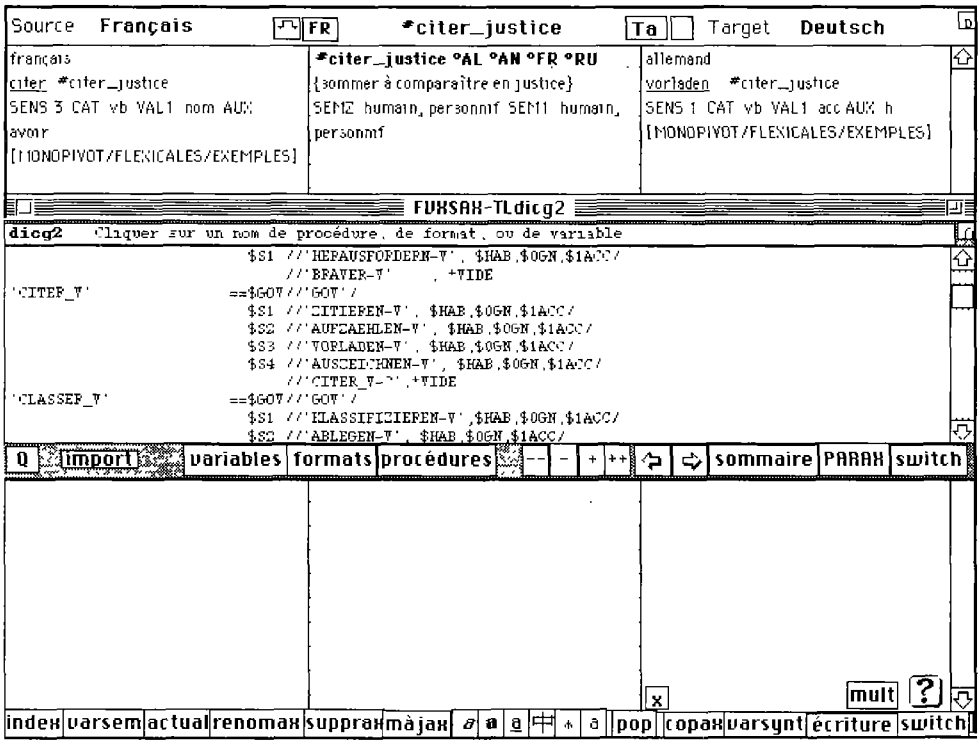

FIGURE 10 Utilisation de PARAX à la maintenance de dictionnaires machine.

7.2. Extraction de dictionnaires bilingues

On peut par ailleurs extraire de PARAX des dictionnaires bilingues sous forme de piles Hypercard avec comme langue source et comme langue cible un couple quelconque des langues représentées dans PARAX.

Ces dictionnaires s'obtiennent simplement à partir de la première carte du dictionnaire monolingue de la langue source (cf. figure 2), en cliquant sur le bouton « bilingue » et en indiquant la langue cible désirée.

Un tel dictionnaire bilingue Hypercard comporte une carte index (analogue à la carte index d'une pile monolingue) permettant d'accéder aux cartes courantes relatives aux lemmes de la langue source. Le format de ces dictionnaires permet de les obtenir sous forme de dictionnaires papier par la simple commande d'impression de la pile.

La figure 11 montre la carte d'un dictionnaire russe-allemand relative au mot allemand *Rittmeister* relevant du passage de sous-acception à sur-acception décrit plus haut. La langue choisie pour la description sémantique peut être arbitrairement sélectionnée parmi les langues représentées dans PARAX, c'est ainsi que l'on a ici un dictionnaire russe-allemand décrivant les sens en français.

russe - allemand	**капитан**	
капитан #captain_officer$_ СМЫСЛ 1 КАТ сущ РОД мр	**#captain_officer$** {officier qui commande une compagnie d'infanterie, un escadron de cavalerie, une batterie d'artillerie *artillerie/air/cavalerie/infanterie*} **#captain_officer$artill** *artillerie* **#captain_officer$air** *air* **#captain_officer$infant** *infanterie* **#captain_officer$caval** { *cavalerie* }	#captain_officer$ Batteriechef #captain_officer$artill SENS 1 CAT nc GNR m Hauptmann #captain_officer$air SENS 1 CAT nc GNR m Hauptmann #captain_officer$infant SENS 2 CAT nc GNR m Rittmeister #captain_officer$caval SENS 1 CAT nc GNR m
капитан #captain_navy СМЫСЛ 2 КАТ сущ РОД мр	**#captain_navy** {officier qui commande un navire de commerce ou de pêche}	Kapitan #captain_navy SENS 1 CAT nc GNR m
капитан #captain_sportteam СМЫСЛ 3 КАТ сущ РОД мр	**#captain_sportteam** {chef d'une equipe sportive}	Mannschaftskapitan #captain_sportteam SENS 2 CAT nc GNR m

FIGURE 11 · Une carte d'un dictionnaire bilingue russe-allemand extrait de PARAX

8. Enrichissement et maintenance de la base

L'enrichissement et la maintenance sont des problèmes délicats dans une base de structure pivot, même de dimensions réduites, du fait de l'interaction entre tous les dictionnaires monolingues par l'intermédiaire du dictionnaire pivot.

Nous n'allons pas décrire ici en détail la méthodologie utilisée pour l'enrichissement et la maintenance de PARAX mais seulement indiquer quelques-uns des outils que nous avons développés.

8.1. Entrée d'une nouvelle acception ; les cartes de description sémantique et le filtrage des acceptions

L'entrée d'une nouvelle acception s'effectue à partir d'un dictionnaire monolingue. En effet, une acception n'est pas définie d'une manière intrinsèque mais a besoin d'au moins un représentant dans une langue pour être définie.

Un problème préliminaire est de connaître avec suffisamment de précision l'état de la base : est-on sûr que cette acception n'a pas déjà été entrée, et si non, comment les éventuelles acceptions voisines ont-elles été définies ?

À cette fin, on peut faire apparaître, dans le champ supérieur droit des cartes courantes des dictionnaires monolingues, une liste des acceptions existantes filtrées suivant la catégorie sémantique. La figure 12 montre, sur une carte en cours de création, la liste des acceptions filtrées par la catégorie sémantique « animal ». Le choix de la

catégorie sémantique de filtrage s'effectue à l'aide de cartes de description sémantique comme celles de la figure 13 : en cliquant sur une catégorie sémantique figurant dans les cases du tableau, on obtient la liste des acceptions correspondant à cette catégorie sémantique et aux catégories hiérarchiquement inférieures.

Enfin, en cliquant sur le nom d'une acception, dans la liste filtrée des acceptions, on obtient sa description, dans le champ inférieur droit, ainsi que la liste des langues où elle est représentée.

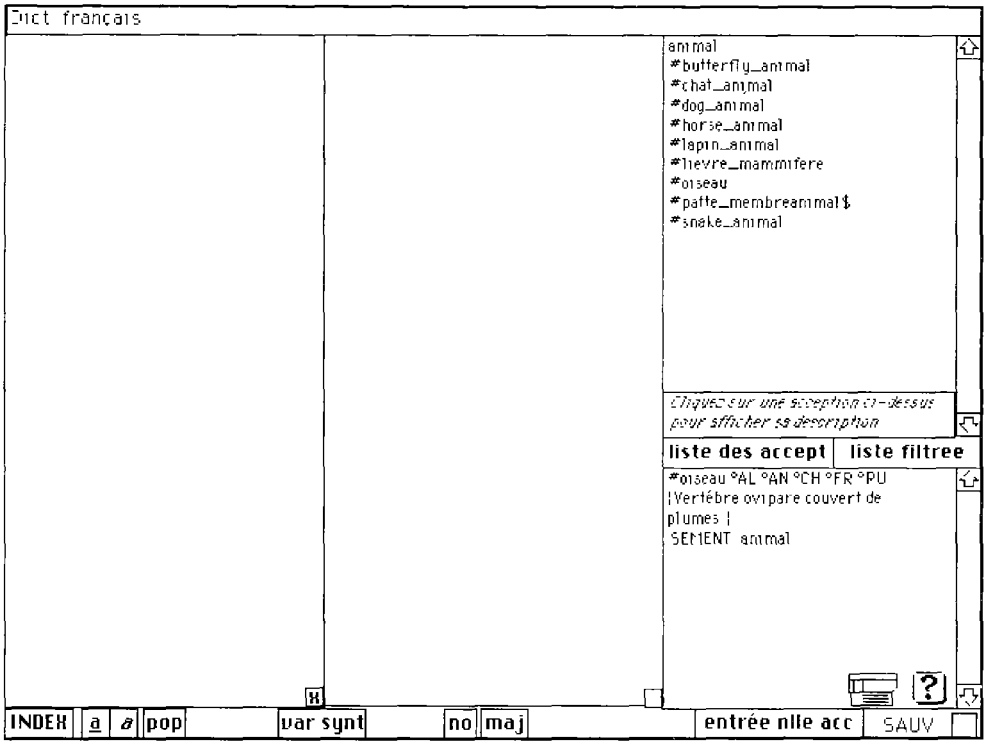

FIGURE 12 : Une carte en cours de création de dictionnaire monolingue comportant une liste filtrée d'acceptions.

8.2. Entrée d'un nouveau lemme

Alors qu'une nouvelle acception est entrée à partir du dictionnaire monolingue de son premier représentant, un nouveau lemme associé à une acception existante est entré à partir du dictionnaire d'acceptions, ce qui permet d'avoir sous les yeux les équivalents déjà entrés dans d'autres langues.

8.3. Écriture contrôlée des variables et valeurs sémantiques et morphosyntaxiques

Les cartes de descriptions sémantiques comme celle de la figure 13, et les cartes de descriptions morphosyntaxiques comme celle représentée pour le français à la figure

15, permettent par ailleurs d'écrire de façon rapide et contrôlée la partie variables-valeurs des descriptions sémantiques et morphosyntaxiques : en cliquant sur le nom des variables et des valeurs désirées, celles-ci s'inscrivent avec une ponctuation normalisée dans le champ inférieur droit de la carte, d'où elles sont ensuite copiées vers les emplacements qui leur reviennent dans le dictionnaire monolingue (variables morphosyntaxiques) ou le dictionnaire d'acceptions (variables sémantiques).

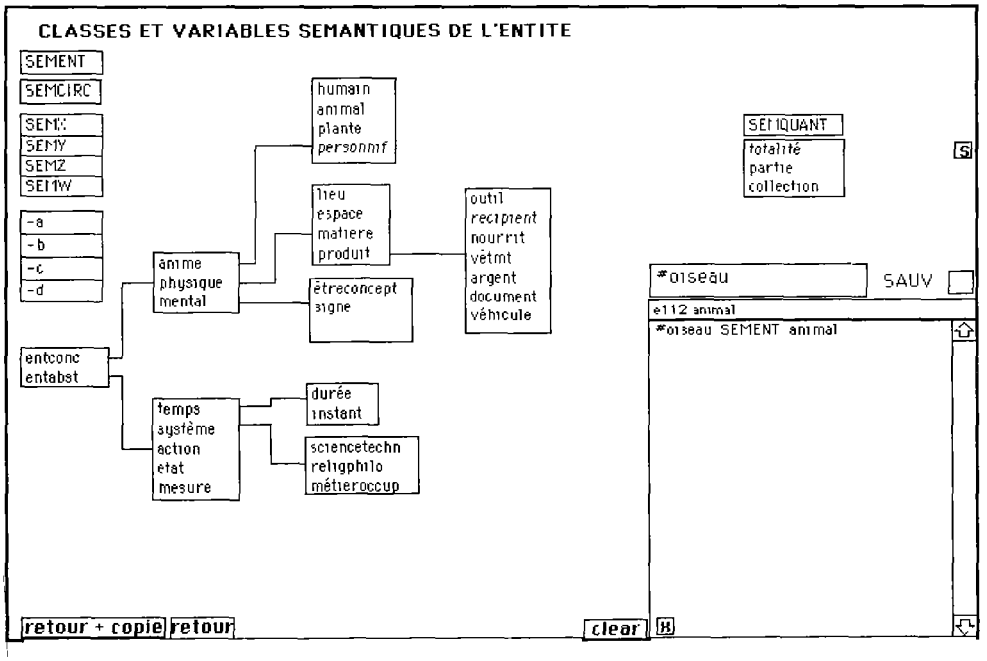

FIGURE 13 . La carte de description sémantique de l'entité.

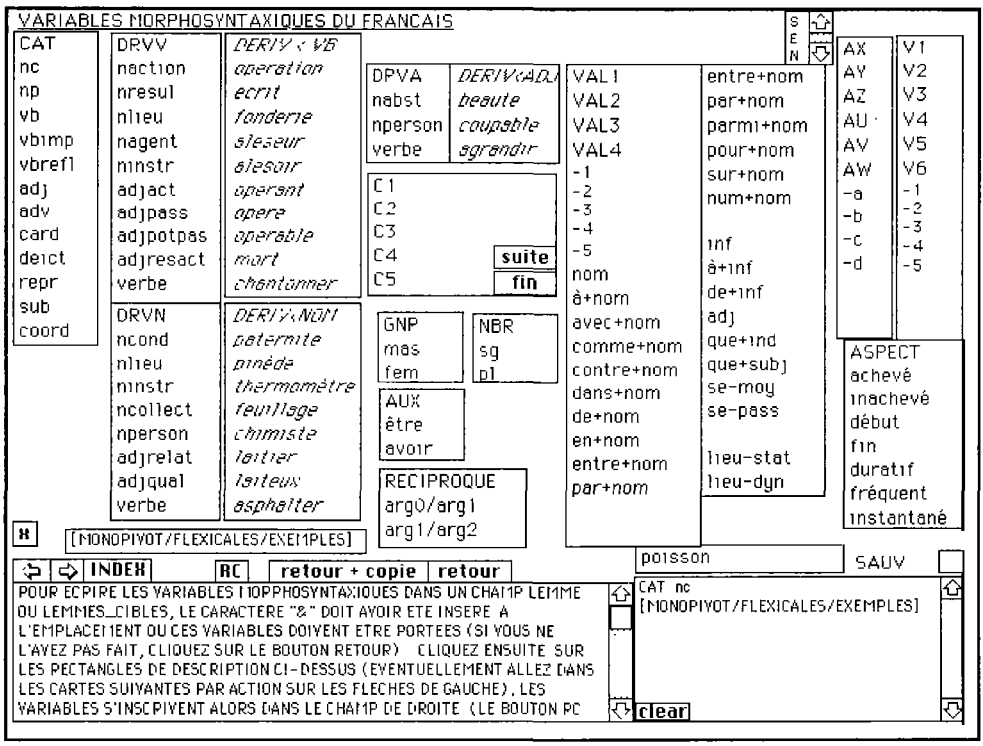

FIGURE 14 : La carte de description morphosyntaxique du français.

8.4. Actualisations diverses

8.4.1. Actualisation des dictionnaires monolingues

Rappelons qu'une carte de pile monolingue contient la description morphosyntaxique « originale » des différents sens d'un lemme donné, mais seulement des copies des descriptions sémantiques des acceptions multilingues correspondantes (cf. 4.2.). Il peut donc être nécessaire de réactualiser ces copies lorsqu'on a été amené à modifier des descriptions sémantiques.

Deux modes de réactualisation sont disponibles :

- dès qu'une modification d'une description d'acception est effectuée dans le dictionnaire d'acceptions, un bouton d'actualisation répercute automatiquement cette modification dans tous les dictionnaires monolingues ;

- par précaution, une procédure de réactualisation, en principe redondante, est disponible dans les piles monolingues : un bouton présent sur chaque carte de lemme permet de remplacer automatiquement les descriptions sémantiques éventuellement périmées par les descriptions présentes dans la pile d'acceptions. Cette réactualisation peut se faire au niveau de la pile entière.

8.4.2. Actualisation du dictionnaire d'acceptions

De même, une carte du dictionnaire d'acceptions contient la description « originale » d'une acception, mais seulement la copie des descriptions morphologiques des représentants de cette acception dans les langues de la base.

Une actualisation de ces informations morphologiques dans le dictionnaire d'acceptions n'est cependant pas vraiment nécessaire, puisqu'elle se réalise automatiquement lors de chaque consultation manuelle (cf. § 4), et que l'extraction automatique de dictionnaires bilingues (cf. § 5) n'utilise que les informations originales des dictionnaires.

Par contre, d'autres fonctionnalités, que nous ne décrirons pas ici, permettent de maintenir la cohérence de la base vis-à-vis de modifications sur les acceptions telles qu'un changement de dénomination de l'acception, une structuration en sous-acceptions, une suppression pure et simple...

Conclusion

La structure de PARAX s'est avérée représenter assez bien le compromis que nous nous étions assignés : assez simple et évolutive pour permettre une expérimentation aisée sur la notion de base multilingue par acceptions, mais disposant de fonctionnalités suffisantes pour que cette expérimentation, qu'il nous reste maintenant à poursuivre, soit réaliste.

Remerciements

Que soient ici remerciés J. P. Guilbaud, N. Nédobejkine et G. Sérasset pour de fructueuses discussions et M. Axtmeyer, D. Levenbach, F. Tchéou et à nouveau N. Nédobejkine pour leur contribution prépondérante à l'entrée des termes anglais, allemands, chinois et russes.

Élaboration d'un dictionnaire informatisé pour le traitement automatique de la langue[1]

Lorne H. Bouchard et Louisette Emirkanian

Département d'informatique et Département de linguistique, Université du Québec à Montréal, Canada

Introduction

Cette recherche s'inscrit dans un programme de recherche dont un des buts est le développement d'une grammaire computationnelle du français (GCF) (Emirkanian & Bouchard, 1992). Cette grammaire, élaborée dans le cadre de la théorie de la Grammaire Syntagmatique Généralisée (GSG) (Gazdar *et al.*, 1985), a été implantée à l'aide de l'outil *Grammar Development Environment* (GDE) (Boguraev, 1988). Du point de vue pratique, une grammaire computationnelle doit avoir une large couverture, si elle veut être utile. D'un point de vue théorique, une grammaire computationnelle doit avoir une couverture large, si elle veut être un modèle crédible de la performance (Bouchard, Emirkanian & Morin, 1992). L'importance du lexique en tant que lieu central de stockage des informations phonologique, morphologique, syntaxique et sémantique est mise de l'avant par la plupart des théories linguistiques contemporaines. Ainsi, une grammaire computationnelle à large couverture doit s'appuyer sur une base de données lexicales importante dans laquelle ces diverses informations sont représentées efficacement. N'ayant trouvé aucune base de données lexicales facilement disponible, nous avons entrepris d'en construire une. La construction d'une telle base de données est une tâche d'envergure qui doit être automatisée le plus possible, ou tout au moins assistée par ordinateur. Nous nous sommes donc concentrés sur le développement de stratégies et d'outils permettant d'extraire les informations nécessaires des dictionnaires sur support informatique et des corpus.

L'ambiguïté, une des caractéristiques des langues naturelles, se manifeste à divers niveaux : nous nous intéresserons ici à l'ambiguïté catégorielle (*le* peut être déterminant ou pronom, par ex.) et à l'ambiguïté structurelle (dans l'analyse de *Paul ex-*

1 Recherche subventionnée par le CRSHC (410-93-0607) et le FCAR (95 ER 1198)

pédie des porcelaines de Chine le syntagme *de Chine* peut être rattaché au syntagme nominal *des porcelaines*, au syntagme verbal *expédie des porcelaines* ou à la phrase comme adjoint de lieu). En faisant intervenir des connaissances – de la langue, du monde et du contexte d'élocution – on peut lever quelquefois cette ambiguïté. L'identification des connaissances requises, leur extraction et leur exploitation est une tâche d'envergure. Néanmoins malgré tout, certaines ambiguïtés subsisteront : leur résolution devra faire intervenir l'humain.

Cependant, puisque certaines informations requises par l'analyseur, la sous-catégorisation et les restrictions de sélection en particulier, ne sont pas notées explicitement dans les dictionnaires disponibles sur support informatique, elles doivent être extraites de l'analyse des exemples utilisés dans les dictionnaires ou de celle de corpus. L'extraction des informations provenant du lexique est un processus itératif qui se développe en plusieurs phases. À chaque étape, une information initiale, déjà disponible ou facile à extraire, est utilisée comme amorce d'un processus qui permet d'extraire de nouvelles informations qui, à leur tour, permettent d'alimenter l'étape suivante.

Interface entre la GCF et le lexique

La GSG est une théorie linguistique qui s'inscrit dans le courant des grammaires à base de contraintes (Shieber, 1986). Dans cette théorie toutes les informations sont véhiculées à l'aide d'ensembles de traits-valeurs, la valeur d'un trait pouvant être un ensemble. Le module d'accès lexical doit produire les ensembles de traits-valeurs correspondant aux mots de la phrase à analyser. Cette tâche comprend essentiellement l'analyse des flexions, la catégorisation et la désambiguïsation catégorielle en contexte s'il y a lieu.

Analyse morphologique

L'analyse morphologique est la première étape de l'analyse de la phrase. Étant donné un mot fléchi, elle produit le radical ou le lemme du mot (c'est-à-dire le mot d'entrée dans le dictionnaire), la catégorie du mot et les informations morphologiques (genre, nombre, personne, temps, mode, par ex.) dérivées de l'analyse de la flexion du mot.

Diverses stratégies peuvent être utilisées pour réaliser des analyseurs morphologiques.

Dictionnaire des formes fléchies

La stratégie la plus simple consiste à utiliser une liste exhaustive des formes fléchies les plus fréquentes. Cette technique a été utilisée dans le module de catégorisation *gramr* de SATO (Daoust, 1992). La liste des formes fléchies a été compilée manuellement en utilisant quelques procédures d'extraction automatique. La liste contient environ 500K mots. La maintenance d'une liste de cette taille pose de sérieux problèmes, d'autant plus que seules les formes fléchies les plus fréquentes sont stockées.

Analyse des suffixes utilisant un dictionnaire des exceptions

Une stratégie originale exploite les régularités de la morphologie du français. La technique est décrite par Meunier (1970). Par exemple, on observe en français que la plupart des mots se terminant en *-able* sont des adjectifs, sauf un petit nombre d'exceptions. Nous avons effectué une étude systématique de la nomenclature du Grand Robert (Rey, 1988). Des 99 134 mots dans la nomenclature, 1 185 se terminent en *-able* et seulement 25 de ces mots correspondent à un nom commun simple ou composé. On trouve également 15 verbes qui ont *-abler* pour suffixe. On peut construire un analyseur morphologique basé sur cette technique, en consultant d'abord un fichier contenant la liste des exceptions à une règle donnée : les mots qui se terminent par *-able* sont des adjectifs. Tous les mots qui n'ont pas été catégorisés lors de la consultation du fichier sont donc catégorisés *adjectif.*

Automate à suffixes et dictionnaire des radicaux

Lors de la construction d'un prototype d'analyseur du français capable de détecter et de corriger certaines erreurs d'orthographe d'usage et de syntaxe (Emirkanian & Bouchard, 1989), nous avions mis au point un analyseur morphologique.

L'algorithme utilisé par cet analyseur morphologique était basé sur un automate à suffixes (Aho & Corasick, 1975). L'analyse de la forme fléchie d'un mot est effectuée de droite à gauche ; lors de ce balayage, diverses hypothèses sont émises, hypothèses qui sont confirmées ou non par une consultation du dictionnaire des radicaux[2]. Dans ce dictionnaire, on associe un radical à la liste des suffixes qui peuvent le compléter. Cette représentation a l'avantage de mettre en facteur la liste des suffixes.

Construction et minimisation d'automates à états finis

Un lexique informatisé a été compilé par le groupe intelligence artificielle et langage naturel du CNET à Lannion (CNET, 1990). Dans ce lexique, un mot est représenté à l'aide de trois informations : le lemme, une liste de suffixes et une liste de paramètres qui définissent la jonction entre le lemme et les suffixes. Fouqueré (1994) a construit un prototype d'analyseur morphologique basé sur ce lexique. Les lemmes sont d'abord développés pour obtenir toutes les formes fléchies, puis un automate à états finis est construit et minimisé. L'analyseur résultant est petit (400K) et très rapide.

Morphologie à deux niveaux : compilateur de règles et de lexique

L'analyse morphologique à deux niveaux (Koskenniemi, 1983) s'inspire de la morphologie concrète : elle postule deux niveaux de représentation, à savoir le niveau lexical et le niveau de surface. La correspondance entre ces deux niveaux est main-

2. Par *radical* on entend dans cette implantation la plus longue partie invariante d'un mot fléchi.

tenue par des transducteurs à états finis bidirectionnels. La morphologie à deux niveaux a été grandement popularisée par un logiciel du domaine public pour l'IBM PC et le Macintosh (Antworth, 1990). L'élégance et la clarté de ce modèle font qu'il a été adopté comme formalisme de choix pour la définition du module d'analyse morphologique de certains systèmes de traitement automatique de la langue naturelle (Ritchie, Russell, Black & Pulman, 1992). En utilisant les techniques de manipulation des automates à états finis, les règles de ce système peuvent être compilées efficacement. Plusieurs travaux portent sur l'analyse morphologique à deux niveaux du français (Karttunen, 1993 ; Karttunen & Beesley, 1992 ; Lun, 1983).

L'analyseur morphologique du français développé par Xerox (Xerox, 1993) utilise cette technologie. Nous précisons que nous avons utilisé cet analyseur comme une boîte noire, puisqu'il nous a été fourni sous la forme d'un programme compilé.

Comparaison des résultats obtenus

On peut comparer sommairement ces divers analyseurs morphologiques[3] en utilisant un fragment d'un corpus d'articles de presse que l'on retrouve à la figure 1. La figure 2 résume les résultats obtenus en utilisant la procédure *gramr* de SATO, l'analyseur XLT de Xerox et le « prototype » d'analyseur construit sur le lexique du CNET. À première vue, il semble y avoir un consensus entre les divers analyseurs, ce qui est encourageant. Cependant une étude plus minutieuse montre que certains mots n'ont pas été analysés, certaines catégories sont absentes et certains mots ont une catégorisation erronée. Par exemple, *n'et aujourd'hui* n'ont pas été analysés par le « prototype » et la catégorie est notée « XXX » dans la figure 2. Les mots *contre* et *grand* ne sont pas notés « Noun » par XLT. La catégorisation de *s'* comme « conjonction » par *gramr* est sûrement une erreur, de même que le grand nombre de « CONJ » produit par le prototype.

Les codes lexicaux utilisés par les analyseurs morphologiques doivent être traduits dans le système de traits-valeurs utilisé par la GCF. Il a été relativement facile de traduire les codes lexicaux produits par les analyseurs morphologiques que nous avons utilisés. En conclusion donc, les analyseurs morphologiques existants peuvent être utilisés avantageusement comme module de pré-traitement de la grammaire GCF.

Désambiguïsation des codes lexicaux en contexte

Quoiqu'on puisse désambiguïser lors de l'analyse syntaxique (Bouchard, Emirkanian & Ratté, 1989), il est préférable de le faire avant, afin de réduire le temps d'analyse. Une solution très répandue consiste à utiliser un algorithme probabiliste qui se fonde sur un modèle de Markov caché (Kupiec, 1992). Cependant, une solution basée sur une exploitation directe des contraintes des langues naturelles (Brill, 1992 ; Chanod & Tapanainen, 1995) nous semble préférable, parce que plus sûre et plus transparente.

3. Il s'agit des analyseurs qui étaient disponibles comme outils autonomes au moment où nous avons effectué notre étude comparative

Nous avons préféré utiliser, comme nous le verrons plus loin, un modèle discret basé sur les n-grammes pour découvrir dans un corpus donné les contraintes fortes qui peuvent servir à désambiguïser les codes lexicaux. Les résultats obtenus semblent meilleurs que ceux obtenus en utilisant un modèle stochastique et surtout plus faciles à justifier, du moins rationnellement.

« Il est difficile de vivre avec des rêves rétrécis Le grand froid qui a figé la fin de l'année 1993 exacerbe l'adversité qui semble aujourd'hui s'acharner contre tous les projets généreux Si nous n'y prenons garde, les temps qui viennent pourraient être ceux de la résignation, ou pire, de la dureté »

Lise Bissonnette, *Le Devoir*,
Vendredi 31 décembre 1993, p A6

FIGURE 1 Fragment d'un article de journal.

mot	gramr	XLT	prototype
a	(aux,v_conj)	(Noun,Verb)	(N,V)
acharner	v_inf	Verb	V
adversité	nomc	Noun	N
année	nomc	Noun	N
aujourd'hui	adv	Adv	XXX
avec	prép	(Adv,Prep)	PREP
ceux	p_indéf	Pro	PRON
contre	(v_conj,nomc,prép)	(Adv,Prep,Verb)	(ADVE,CONJ,N,PREP,V)
de	prép	(Det,Prep)	PREP
des	(artind,artpart,prép)	(Det,Prep+Det)	(CONJ,DET)
difficile	adj	Adj	A
dureté	nomc	Noun	N
est	(aux,v_conj,nomc)	(Noun,Verb)	(A,N,V)
être	(aux,v_inf,nomc)	(Noun,Verb)	(N,V)
exacerbe	v_conj	Verb	V
figé	(adj,ppassé)	(Adj,Verb)	(A,V)
fin	(adj,nomc)	(Adj,Adv,Noun)	(A,ADVE,N)
froid	(adj,nomc)	(Adj,Noun)	(A,N)
garde	(v_conj,nomc)	(Noun,Verb)	(N,V)
généreux	adj	Adj	A
grand	(adj,nomc)	(Adj,Adv)	(A,ADVE,N)
il	p_pers	PC	(CONJ,PRON)
l'	(artdéf,p_pers)	(Det,PC)	(DET,PRON)
la	(artdéf,nomc,p_pers)	(Det,Noun,PC)	(DET,N,PRON)
le	(artdéf,p_pers)	(Det,PC)	(DET,PRON)
les	(artdéf,p_pers)	(Det,PC)	(DET,PRON)
n'	adv	Neg	XXX
nous	p_pers	(PC,Pro)	PRON
ou	conjonction	Coord	CONJ
pire	adj	(Adj,Noun)	(A,N)
pourraient	v_conj	Verb	V
prenons	v_conj	Verb	V
projets	nomc	Noun	N

qui	p_relatif	Pro	PRON
résignation	nomc	Noun	N
rétrécis	(adj,v_conj,ppassé)	Verb	(A,V)
rêves	(v_conj,nomc)	(Noun,Verb)	(N,V)
s'	conjonction	(PC,SConj)	PRON
semble	v_conj	Verb	V
si	(adv,conjonction,nomc)	(APrMod,Noun,SConj)	ADVE,CONJ,N,PREP)
temps	nomc	Noun	(N,V)
tous	(détindéf,p_indéf)	(Det,Pro)	(A,DET,PRON)
viennent	v_conj	Verb	V
vivre	(v_inf,nomc)	(Noun,Verb)	(N,V)
y	(nomc,p_pers)	(Noun,PC)	(N,PRON)
1993	détnum	Card	XXX

FIGURE 2 · Liste de codes lexicaux affectés.

Le problème du rattachement des syntagmes prépositionnels

L'ambiguïté du rattachement des syntagmes prépositionnels donne lieu à une explosion combinatoire dans le nombre des analyses produites par un analyseur (Church & Patil, 1982).

Par exemple, lorsqu'il apparaît dans la position immédiatement à droite du syntagme nominal objet direct du verbe, le syntagme prépositionnel peut être rattaché à trois endroits différents : comme adjoint du syntagme nominal qui le précède immédiatement, comme argument du verbe (le plus souvent facultatif) ou comme adjoint du syntagme verbal. À défaut d'autres informations, l'analyseur doit produire trois analyses dans ce cas.

Trois types d'information permettent de modéliser les préférences au niveau du rattachement (Wilks, Huang & Fass, 1985) : la sous-catégorisation, les restrictions de sélection et certaines connaissances du domaine. La sous-catégorisation, une information de nature syntaxique, informe sur la structure des compléments, obligatoires ou facultatifs, d'une tête lexicale. Les restrictions de sélection, une information de nature sémantique, informent sur les types sémantiques des compléments de la tête. Les connaissances du domaine sont des connaissances de nature encyclopédique qui établissent des liens entre les éléments du domaine.

Sous-catégorisation dans GSG et la GCF

Les règles lexicales de dominance immédiate sont la composante de la GSG qui établit le pont entre la grammaire et le lexique. Dans ces règles, la valeur du trait SUB de la tête lexicale représente de façon succincte les divers compléments obligatoires ou facultatifs de la tête. Nous ne traiterons, dans les exemples qui suivent, que de la sous-catégorisation des verbes.

La sous-catégorisation des verbes a été traitée abondamment dans la GCF (GIREIL, 1994). Le verbe *penser* (figure 3) porte les traits de sous-catégorisation SUB20+, SUB33+, SUB50+ et SUB51+.

Traits	Schémas associés	Exemples
SUB20	P3[à,CPR N3]	Gilles pense à Mireille
SUB33	P3[à,CPR V3]	Gilles pense à visiter Boston
SUB50	Q3[que]	Mireille pense que cette thèse est très bonne
SUB51	V3[VFORM er]	Mireille pense venir

FIGURE 3 . Schéma de sous-catégorisation pour *penser*.

Dans le cas d'une sous-catégorisation stricte, c'est-à-dire lorsqu'un complément est obligatoirement requis par la tête, le trait de sous-catégorisation permet d'obtenir une seule analyse syntaxique. Le plus souvent cependant, l'analyseur est confronté à des cas de sous-catégorisation où le complément est spécifié comme facultatif.

Détermination des cas de sous-catégorisation

S'il est fréquent que les dictionnaires courants notent systématiquement la distinction transitif / intransitif pour les verbes, la sous-catégorisation apparaît peu ou le plus souvent elle n'apparaît pas du tout. Les cas de sous-catégorisation doivent être extraits de l'analyse des exemples contenus dans le dictionnaire ou d'une analyse de corpus.

Analyse des exemples d'un dictionnaire

Le ZYZOMYS (ZYZOMYS, 1989) est la version CD-ROM du *Dictionnaire de notre temps* (Guerard, 1989). Sur le CD-ROM, un fichier contient la base textuelle du dictionnaire.

On distingue quatre sources d'information pour les verbes dans le ZYZOMYS : la catégorie lexicale et le modèle de conjugaison spécifiés dans l'entrée du verbe, les informations entre parenthèses, et les exemples. Pour extraire l'information, nous avons adopté une représentation commune permettant d'unifier en un tout cohérent les informations provenant de diverses sources.

Dans la plupart des cas (90 %), un verbe est spécifié comme *transitif, intransitif, impersonnel* ou *pronominal*. Dans seulement 10 % des cas il est noté *transitif direct, transitif indirect, copule* ou *auxiliaire*. Le ZYZOMYS utilise 83 modèles de conjugaison. En consultant le modèle du verbe on peut retrouver divers renseignements : l'auxiliaire requis, *avoir* ou *être* ou les deux, et si le verbe est *pronominal* ou *défectif*.

Le ZYZOMYS utilise les parenthèses dans les articles pour signaler des restrictions de sélection sur les arguments des verbes.

Par exemple, dans l'article de abandonner, on trouve les informations suivantes :

```
renoncer à (qqch)
laisser (qqch) à (qqn)
mettre (qqch) à la disposition de (qqn)
délaisser (qqch)
```

Cependant, une analyse systématique des informations signalées entre parenthèses dans les entrées révèle que cette information est très éparse, comme l'indique l'affichage en ordre inverse des fréquences à la figure 4. Nous n'avons retenu de cette information que la distinction *qqn/qqch* et nous avons encodé toutes les autres distinctions en termes de celle-là.

```
342   (qqn)
164   (qqch)
 87   (S. comp.)
 50   (choses)
 37   (personnes)
 34   (e)
 27   (+ inf.)
 17   (une chose)
 15   (I)
 13   (sens 2)
 12   (un navire)
 12   (qqn, qqch)
 12   (Re'cipr.)
 12   (E)
 11   (un corps)
 11   (un animal)
 11   (Personnes.)
 11   (Choses.)
 10   (un objet)
 10   (un liquide)
 10   (un lieu)
...
```

FIGURE 4 · Informations notées entre parenthèses dans les articles du ZYZOMYS.

Comme dernière source d'information sur les verbes dans le ZYZOMYS, nous avons les exemples utilisés dans les articles. Ces exemples sont simples et l'information qu'on peut en extraire, les cas de sous-catégorisation, permet d'enrichir les informations extraites par ailleurs.

À titre indicatif, voici quelques exemples utilisés pour illustrer l'article d'*abandonner* :

```
Abandonner un projet.
Abandonner un emploi.
De nombreux coureurs ont abandonné au cours de cette étape.
J'abandonne la capitale pour m'établir dans une petite ville.
Abandonner sa famille.
```

Pour analyser ces exemples nous avons développé une grammaire hors-contexte empirique permettant d'isoler les syntagmes qui gravitent autour du verbe. Un « chart parser » piloté par cette grammaire permet d'analyser les exemples.

Les articles des verbes sont d'abord extraits du dictionnaire et traduits en format LISP à l'aide d'un programme **lex**. Un lexique des formes fléchies uniques est construit à partir de ces données et catégorisé à la main. Les exemples sont ensuite analysés à l'aide du « chart parser » et les données en format LISP sont enrichies du résultat de l'analyse syntaxique. La figure 5 montre un fragment du résultat correspondant à l'analyse du verbe *abandonner*.

```
'(ABANDONNER (CAT V)
        ((SENS 1
                ((SCAT TRANSITIF (TAKES SN))
                 (AUX AVOIR)
                 (SUBDEF 1
                        ((SPECIFIEUR (ANIME -))
                         (EXEMPLE (TAKES SN)
                                (ABANDONNER UN
                                        PROJET))
                         (EXEMPLE (TAKES SN)
                                (ABANDONNER SON
                                        EMPLOI))
                         (DOMAINE SPORT
                                ((TAKES 0)
                                 (EXEMPLE AMBIGUE
                                        (DE NOMBREUX
                                           COUREURS
                                           ONT
                                           ABANDONNE03
                                           AU
                                           COURS
                                           DE
                                           CETTE
                                           E03TAPE))))))
                 (SUBDEF 2
                        ((TAKES
                          (SN (ANIME -)
                             SP
                             ((HUMAIN +) (PREP A))))
                         (SPECIFIEUR (ANIME -))
                         (SPECIFIEUR (HUMAIN +))))
                 (SUBDEF 3
                        ((SPECIFIEUR (ANIME -))))
                 (SUBDEF 4
                        ((EXEMPLE AMBIGUE
                                (J ABANDONNE
                                   LA
                                   CAPITALE
                                   POUR
                                   M
                                   E03TABLIR
                                   DANS
                                   UNE
                                   PETITE
                                   VILLE))))
                 (SUBDEF 5
                        ((EXEMPLE (TAKES SN)
                                (ABANDONNER SA
                                        FAMILLE))))))
...
```

FIGURE 5 Fragment de l'information extraite du ZYZOMYS pour le verbe *abandonner*.

La figure 6 permet d'apprécier le taux de réussite d'une première expérience que nous avons effectuée. On peut observer que la transcription des entrées du dictionnaire en formes LISP ne s'effectue pas toujours correctement à cause d'imperfections dans le programme de transcription. On peut voir également que les échecs de l'analyse des exemples sont attribués, à part égale, à deux causes principales : les ambiguïtés et les erreurs d'analyse. L'ambiguïté provient des analyses mutiples, dues notamment au problème du rattachement. Les erreurs d'analyse sont principalement dues aux limitations de la grammaire empirique.

entrées	nombre	« lispifiées »	exemples	corrects	ambiguës	erronés
a-	411	387 (94%)	764 (100%)	309 (40%)	222 (30%)	233 (30%)
b-	236	125 (52%)	87 (100%)	44 (50%)	19 (22%)	24 (28%)
c-	403	299 (74%)	409 (100%)	172 (42%)	116 (28%)	121 (30%)
d-	776	687 (89%)	876 (100%)	406 (46%)	243 (28%)	229 (26%)
e-	641	599 (93%)	884 (100%)	407 (46%)	272 (31%)	205 (23%)
Total	2467	2097 (85%)	3020 (100%)	1338 (44%)	872 (29%)	812 (27%)

FIGURE 6 Statistiques sur le taux de réussite de l'extraction des exemples du ZYZOMYS

L'analyse du ZYZOMYS a été effectuée en utilisant un « chart parser » piloté par une grammaire hors-contexte façonnée à la main. Dans des travaux subséquents, nous avons cherché à éliminer cette étape dans la mesure où elle est très laborieuse et dépendante du corpus. Une première direction de recherche a exploré l'induction automatique de grammaires régulières, tandis que la seconde a étudié l'exploitation directe d'un modèle discret des n-grammes, ce qui élimine complètement la grammaire.

Analyse d'un corpus

Le corpus d'Alain Guillet (1990) est un ensemble de phrases simples permettant d'exemplifier la classification des verbes dans le lexique-grammaire (Leclère, 1990). Ce corpus contient 20 603 phrases et 175 021 mots, dont 17 503 formes fléchies uniques.

Considérons la première direction.

Induction de grammaire régulière

Dans une première expérience nous avons cherché à induire automatiquement une grammaire régulière (Ejerhed, 1988) à partir d'un échantillon du texte à analyser. Cela revient à effectuer une induction à partir de données positives seulement et certaines précautions sont nécessaires afin d'éviter la sur-généralisation (Angluin, 1980). Pour cela, il suffit d'ordonner les données présentées pour que le processus d'induction converge sur le langage le plus restreint qui soit compatible avec les données, stratégie connue sous le nom de principe des sous-ensembles (Berwick, 1986). L'ordonnancement requis consiste à toujours présenter les phrases et les syntagmes plus

simples avant les plus complexes. Nous avons utilisé comme procédure d'induction la méthode de fusion des queues (Miclet, 1980) qui est particulièrement simple à implanter.

Manuellement, un échantillon du corpus a été catégorisé, découpé en constituants et ordonné selon le principe des sous-ensembles. Puisque l'algorithme semblait sur-généraliser, du moins par rapport au type de grammaire recherché, nous avons dû noter par un & les constituants qui étaient vides dans les phrases utilisées pour l'apprentissage et modifier l'algorithme d'induction en conséquence.

La figure 7 illustre la séquence d'apprentissage utilisée pour induire une grammaire de phrases simples qui découpe les constituants qui gravitent autour du verbe, par exemple le sujet, le complément direct, le complément indirect. Les | servent à délimiter les syntagmes superficiels de la phrase et les &, qui marquent des positions vides, servent à guider l'induction dans une position donnée.

```
| &       | V | &       | &       |.
| NP      | V | &       | &       |.
| DET N   | V | &       | &       |.
| &       | V | NP      | &       |.
| &       | V | DET N   | &       |.
| &       | V | &       | P NP    |.
| &       | V | &       | P DET N |.
...
```

FIGURE 7 : Une séquence d'apprentissage bien ordonnée.

L'automate à états-finis résultant est converti manuellement en un transducteur qui permet de découper le reste du corpus en constituants. Ce transducteur est donc un analyseur syntaxique primitif qui permet de mettre en relief les arguments du verbe.

Afin de cerner le positionnement de la négation et des clitiques autour du verbe conjugué simple, on doit avoir recours à une séquence d'apprentissage comme le montre la figure 8. Dans cette séquence et la suivante, le signe « * » identifie un mot absent et le signe « & » , un constituant vide.

```
&  | * * * V *                  | &  | &  | &.
&  | * * * V ADV *              | &  | &  | &.
&  | * * CLI_I V *              | &  | &  | &.
&  | NEG1 * * V NEG2            | &  | &  | &.
&  | NEG1 * * V ADV NEG2        | &  | &  | &.
&  | NEG1 CLI_D * V NEG2        | &  | &  | &.
&  | NEG1 * CLI_I V NEG2        | &  | &  | &.
...
```

FIGURE 8 : Séquence d'apprentissage pour le verbe avec positionnement des clitiques et de la négation.

Afin de cerner le positionnement de la négation et des clitiques autour du verbe conjugué à un temps composé, on doit avoir recours à une séquence d'apprentisage comme celle de la figure 9.

```
&  | NEG1 CLI_D * AUX NEG2 P_PAS  | & | & | &.
&  | NEG1 * CLI_I AUX NEG2 P_PAS  | & | & | &.
&  | NEG1 * * AUX ADV NEG2 P_PAS  | & | & | &.
&  | * CLI_D * AUX * P_PAS        | & | & | &.
&  | * * CLI_I AUX * P_PAS        | & | & | &.
&  | * * * AUX ADV * P_PAS        | & | & | &.
...
```

FIGURE 9 Séquence d'apprentissage pour l'auxiliaire et le participe passé avec positionnement des clitiques et de la négation

Cette méthode présente deux désavantages majeurs. En effet, le premier est que les données doivent être prétraitées manuellement, ce qui demande un travail fastidieux. Si on cherche à induire une grammaire particulière, le travail de préparation des séquences d'apprentissage devient rapidement comparable à celui qui est requis pour synthétiser manuellement la grammaire désirée. Le second désavantage est qu'il est difficile de prévoir la séquence d'entraînement nécessaire pour obtenir une grammaire qui couvre un corpus donné.

De plus, malgré les précautions prises, l'algorithme sur-généralise et il en résulte que de nombreuses phrases non-grammaticales sont acceptées par la grammaire. Cette limitation a peu d'impact dans notre application. Cependant, nous avons observé que les grammaires obtenues en désactivant l'étape d'induction demeuraient de taille acceptable et sur-généralisaient moins. L'induction présente donc peu d'intérêt en pratique pour ce genre d'application, ce qui nous amène à exposer la deuxième direction de recherche que nous avons explorée.

Désambiguïsation catégorielle en contexte

Dans cette seconde approche nous avons expérimenté avec un modèle discret des n-grammes, c'est-à-dire ne faisant aucunement intervenir les notions de probabilité ou de statistique. Plutôt que de réitérer l'expérience précédente et de catégoriser manuellement, nous avons préféré utiliser un programme pour le faire. Cela nous permettra de discuter à nouveau du problème de la désambiguïsation et de proposer une nouvelle solution.

Un lexique des formes fléchies uniques extraites du corpus a été catégorisé en utilisant le tagger XLT de Xerox (Xerox, 1993) afin d'obtenir pour chaque forme fléchie sa classe d'ambiguïté, c'est-à-dire l'ensemble des étiquettes de catégorie qui peuvent la caractériser. On peut voir affichées à la figure 10 les classes d'ambiguïté correspondant aux premières entrées dans le lexique du corpus. Puis nous avons étiqueté le corpus directement avec une version expérimentale du tagger stochastique de Xerox (Kupiec, 1992 ; Xerox, 1995).

```
A NOUN_PL NOUN_SG PREP_A
a NOUN_PL NOUN_SG VAUX_P3SG
à PREP_A
a NOUN_PL NOUN_SG VAUX_P3SG
à PREP_A
a-t-elle PC+VAUX_P3SG
a-t-il PC+VAUX_P3SG
abaissé PAP_SG
abaissement NOUN_SG
abandon NOUN_SG
abandonne VERB_P1P2 VERB_P3SG
abandonné PAP_SG
abandonnée PAP_SG
abandonnent VERB_P3PL
abasourdi PAP_SG
abasourdissement NOUN_SG
abat NOUN_SG VERB_P3SG
abat-jour NOUN_PL NOUN_SG
...
```

FIGURE 10 Fragment du lexique des formes fléchies et des classes d'ambiguité correspondantes.

Nous avons dû corriger manuellement un grand nombre d'erreurs d'étiquetage : *Cela/VERB_P3SG, Le/DET_SG, maçon/ADJ_SG, ...d'/DET_PL être/VAUX_INF,* par exemple. Il semble que ce tagger stochastique ne tire pas toujours parti du contexte dans lequel se trouve le mot. Lors de cette phase d'édition, nous avons également ajouté quelques distinctions supplémentaires puisque Xerox ne distingue que PREP_A et PREP_DE et regroupe les autres prépositions simplement sous PREP alors qu'elles sont souvent syntaxiquement discriminantes. À la figure 11 on peut voir un fragment du corpus étiqueté et désambiguïsé.

```
Max/NOUN_PRP a/VAUX_P3SG abaissé/PAP_SG Léa/NOUN_PRP à/PREP_A
ce/PRON qu'/CONJQUE elle/PRON demande/VERB_P3SG 100/NUM
f./NOUN_INV ./SENT

Max/NOUN_PRP a/VAUX_P3SG abaissé/PAP_SG Léa/NOUN_PRP
à/PREP_A demander/VERB_INF 100/NUM f./NOUN_INV ./SENT

Max/NOUN_PRP s'/PC est/VAUX_P3SG abaissé/PAP_SG à/PREP_A
demander/VERB_INF 100/NUM f./NOUN_INV ./SENT

Le/DET_SG vent/NOUN_SG a/VAUX_P3SG
abaissé/PAP_SG la/DET_SG température/NOUN_SG ./SENT

Le/DET_SG vent/NOUN_SG a/VAUX_P3SG causé/PAP_SG l'/DET_SG
abaissement/NOUN_SG de/PREP_DE la/DET_SG température/NOUN_SG ./SENT
...
```

FIGURE 11 : Fragment du corpus étiqueté

L'algorithme expérimental basé sur un modèle discret des n-grammes fonctionne en deux étapes. Dans la phase d'apprentissage on mémorise tous les n-grammes construits sur les catégories désambiguïsées des phrases du corpus d'entraînement. Dans la phase d'exploitation, on ne tient plus compte des catégories désambiguïsées des phrases du corpus, on consulte plutôt le lexique qui retourne la classe d'ambiguïté d'une forme fléchie et on utilise les n-grammes mémorisés lors de la phase d'apprentissage comme contraintes permettant de désambiguïser les catégories des mots de la phrase. Les résultats préliminaires montrent qu'en utilisant seulement 10 % du corpus étiqueté comme apprentissage, on obtient un taux de désambiguïsation exact à plus de 87 %. Il semble que des erreurs d'étiquetage soient la cause de ce taux de succès encore trop faible.

Segmentation de la phrase en constituants

Une fois les catégories lexicales désambiguïsées en contexte, il est facile de découper les phrases du corpus en constituants de premier niveau. L'algorithme le plus simple consiste à balayer la phrase de gauche à droite à la recherche d'un verbe conjugué, le pivot de la phrase. À partir de ce point, le segment correspondant au noyau verbal est construit en effectuant un balayage à gauche pour récupérer les clitiques, les adverbes et la négation, puis un balayage à droite pour récupérer le participe passé et les adverbes. Ensuite, on construit des segments avec le fragment de phrase qui précède le segment noyau verbal et le fragment de phrase qui le suit, borné[4] par une préposition, s'il y a lieu. À partir de ce point dans la phrase, on construit itérativement des segments prépositionnels qui sont délimités[5] à gauche et bornés à droite par des prépositions ou la fin de phrase, s'il y a lieu. La figure 12 illustre la segmentation en constituants des premières phrases du corpus.

```
1-  [Max][a abaissé][Léa][à ce qu' elle demande 100 f.]
2-  [Max][a abaissé][Léa][à demander 100 f.]
3-  [Max][s' est abaissé][à demander 100 f.]
4-  [Le vent][a abaissé][la température]
5-  [Le vent][a causé][l' abaissement][de la température]
6-  [Max][a abaissé][la toise][sur la tête][de Luc]
7-  [Max][a abaissé][la manette]
8-  [Paul][a abandonné]
...
```

FIGURE 12 : Illustration de la segmentation des phrases du corpus.

Ad hoc certes, cette méthode d'analyse a l'avantage d'être simple, robuste et facile à utiliser, du moins sur les corpus de phrases simples qui nous intéressent.

On peut constater que cette segmentation étale bien les arguments du verbe. L'algorithme de segmentation est donc une technique qui permet de mettre en relief

4 La préposition ne fait pas partie du segment en question
5 La préposition fait partie du segment en question

les cas de sous-catégorisation d'un verbe. Cependant, on observe que les **P2** qui modifient le **N2** objet direct du verbe ou qui modifie le **N2** qui est la tête du **P2** sont étalés comme si c'étaient des arguments du verbe (voir les exemples 5 et 6 de la figure 12). Tous les exemples de ce type devront être exclus d'un corpus utilisé pour l'apprentissage des cas de sous-catégorisation.

L'extraction des têtes des segments s'effectue par un simple balayage de gauche à droite de ces segments avec certains petits ajustements qui permettent de traiter des complétives et des infinitives. La figure 13 illustre les têtes extraites des exemples segmentés présentés ci-dessus.

Nous pouvons maintenant reproduire les résultats obtenus lors de l'analyse du ZYZOMYS, mais beaucoup plus rapidement.

Cependant, le problème du rattachement des **P2** demeure intégral et pour le résoudre nous devons faire appel aux restrictions de sélection, qui doivent être extraites de l'analyse d'exemples de dictionnaire ou de celle de corpus. Nous avons progressé néanmoins dans la mesure où nous avons un découpage brut des constituants.

```
[("Max" NOUN_PRP)]
[("a" VAUX_P3SG) ("abaissé" PAP_SG)]
[("Léa" NOUN_PRP)]
[("à ce qu'" PREP_A) ("elle demande 100 f." COMPL]

[("Max" NOUN_PRP)]
[("a" VAUX_P3SG) ("abaissé" PAP_SG)]
[("Léa" NOUN_PRP)]
[("à" PREP_A) ("demander 100 f." INF)]

[("Max" NOUN_PRP)]
[("s'" PC) ("est" VAUX_P3SG) ("abaissé" PAP_SG)]
[("à" PREP_A)("demander 100 f." INF)]

[("vent" NOUN_SG)]
[("a" VAUX_P3SG) ("abaissé" PAP_SG)]
[("température" NOUN_SG)]

[("vent" NOUN_SG)]
[("a" VAUX_P3SG) ("causé" PAP_SG)]
[("abaissement" NOUN_SG)]
[("de" PREP_DE) ("température" NOUN_SG)]

[("Max" NOUN_PRP)]
[("a" VAUX_P3SG) ("abaissé" PAP_SG)]
[("toise" NOUN_SG)]
[("sur" PREP) ("tête" NOUN_SG)]
[("de" PREP_DE) ("Luc" NOUN_PRP)]

[("Max" NOUN_PRP)]
[("a" VAUX_P3SG) ("abaissé" PAP_SG)]
[("manette" NOUN_SG)]

[("Paul" NOUN_PRP)]
[("a" VAUX_P3SG) ("abandonné" PAP_SG)]
...
```

FIGURE 13 : Illustration de l'extraction des têtes des segments.

Détermination des restrictions de sélection

Les restrictions de sélection sont les contraintes sémantiques qui portent sur les arguments d'une tête lexicale qui permettent de juger de la bonne formation des syntagmes construits sur ces têtes. Pour spécifier les restrictions de sélection, nous utilisons une classification de l'espace des noms qui est basée sur une structure appelée ontologie.

Choix d'une ontologie

Construite sur un petit nombre de distinctions élémentaires, une ontologie permet de discerner les classes de noms significatives du point de vue des restrictions de sélection. Dans la perspective des systèmes de types utilisés en programmation fonctionnelle ou dans la théorie *Head-Driven Phrase Structure Grammar* (HPSG) (Pollard & Sag, 1987), ces classes de noms correspondent à des types complexes définis par héritage multiple. Ces liens d'héritage permettent de généraliser facilement à partir d'un concept donné. Cette ontologie n'est pas une taxinomie dans la mesure où une classe peut être définie à partir de plusieurs classes existantes. La structure de l'ontologie prend la forme d'un treillis plutôt que celle d'une arborescence. Dans la mesure où le principe de classification utilisé fait intervenir des connaissances naïves de la réalité, c'est-à-dire du monde, cette structure de classification mérite d'être appelée une ontologie.

Nous avons effectué une comparaison de trois ontologies qui ont été proposées pour l'analyse automatique du langage naturel, les ontologies des projets CORE Language Engine et PENMAN ainsi que celle développée par Dahlgren. L'ontologie PENMAN a été développée dans le cadre d'un projet (Bateman, 1991 et 1993) visant à développer un système général de génération de textes en langue naturelle dans le paradigme des grammaires systémiques. L'ontologie développée pour le CORE Language Engine par SRI à Cambridge (Alshawi, 1992) a pour fonction la désambiguïsation syntaxique et sémantique dans un système de traitement automatique. L'ontologie qui est présentée par Dahlgren (1988 et 1993) a été développée pour un système de compréhension automatique de textes en langue naturelle. Si notre étude a pu mettre en relief de nombreuses similitudes entre ces ontologies, elle a aussi permis d'identifier plusieurs points de divergence. Nous avons adopté, de façon provisoire, l'ontologie de Dahlgren qui est reproduite à la figure 14.

ENTITY	(ABSTRACT | REAL) & (INDIVIDUAL | COLLECTIVE)
ABSTRACT	IDEAL | PROPOSITIONAL | QUANTITY | IRREAL
QUANTITY	NUMERICAL | MEASURE
REAL	(PHYSICAL | TEMPORAL | SENTIENT) & (NATURAL | SOCIAL)
PHYSICAL	(STATIONARY | NONSTATIONARY) & (LIVING | NONLIVING)
NONSTATIONARY	(SELFMOVING | NONSELFMOVING)
COLLECTIVE	MASS | SET | STRUCTURE
TEMPORAL	RELATIONAL | NONRELATIONAL
RELATIONAL	(EVENT | STATE) & (MENTAL | EMOTIONAL | NONMENTAL)
EVENT	(GOAL | NONGOAL) & (ACTIVITY | ACCOMPLISHMENT | ACHIEVEMENT)

FIGURE 14 Ontologie tirée de Dahlgren & McDowell (1986).

Extraction des restrictions de sélection

Le type associé à la tête nominale d'un syntagme est obtenu en classifiant celle-ci selon l'ontologie, une étape qui exige pour le moment une intervention de l'utilisateur.

Les types des arguments d'un même verbe sont unifiés pour obtenir un type résultant. En général, ce type résultant correspond au plus petit des majorants des types des arguments, cependant il y a lieu de dédoubler parfois la restriction de sélection, dans le cas où le type résultant est trop général par exemple.

Examinons les exemples du verbe abaisser dans le corpus (figure 13). On retrouve « Max » ou « vent » comme têtes des segments sujets et « manette », « température » et « toise » comme têtes des segments objets directs. L'ontologie classifie ces noms de la façon suivante :

```
Max          (ENTITY (REAL SENTIENT NATURAL) INDIVIDUAL)

Max          (ENTITY (REAL (PHYSICAL (NONSTATIONARY SELFMOVING)
              LIVING) NATURAL) INDIVIDUAL)

vent         (ENTITY (REAL (PHYSICAL (NONSTATIONARY SELFMOVING)
              NONLIVING) NATURAL) COLLECTIVE)

manette      (ENTITY (REAL (PHYSICAL (NONSTATIONARY NONSELFMOVING)
              NONLIVING) SOCIAL) INDIVIDUAL)

température  (ENTITY (ABSTRACT (QUANTITY MEASURE)) INDIVIDUAL)

toise        (ENTITY (REAL (PHYSICAL (NONSTATIONARY NONSELFMOVING)
              NONLIVING) SOCIAL) INDIVIDUAL)
```

L'unification de la première entrée pour « Max » (de type PERSON) et de l'entrée pour « vent » (de type FLOW_GROUP) produit le type « (ENTITY (REAL NATURAL)) » tandis que l'unification de la seconde entrée pour « Max » (de type HUMAN) et de l'entrée pour « vent » (de type FLOW_GROUP) produit le type « (ENTITY (REAL (PHYSICAL (NONSTATIONARY SELFMOVING)) NATURAL)) ». On retient donc ce second type puisqu'il est plus spécifique que le premier.

L'unification du type de « manette » (MANMADEOBJ) et de celui de « toise » (MANMADEOBJ) produit le type « (ENTITY (REAL (PHYSICAL (NONSTATIONARY NONSELFMOVING) NONLIVING) SOCIAL) INDIVIDUAL) », c'est-à-dire le même type que celui des deux arguments. Ce type caractérise, en partie du moins, le sens 1 de ce verbe dans le Grand Robert :

> ◊ 1. Faire descendre à un niveau plus bas ⇒ **Baisser** *Voulez-vous abaisser la vitre ? Abaisser qqch. en inclinant*, en penchant*.* — Arithm *Abaisser un chiffre* dans une division, écrire un chiffre du dividende à la suite du reste obtenu — Géom. *Abaisser une perpendiculaire* : mener d'un point une perpendiculaire à une ligne, à un plan.

Par contre, l'unification du type de « manette » ou de celui de « toise » avec celui de « température » (MEASUREMENT) produit le résultat « (ENTITY INDIVIDUAL) », c'est-à-dire un des types les plus vagues dans l'ontologie. Ce résultat signale qu'on doit prévoir une seconde entrée pour les restrictions de sélection du verbe abaisser qui correspond au sens 3 de ce verbe dans le Grand Robert :

> ◊ 3. Diminuer la quantité de, faire baisser. *Abaisser la température* ⇒ **Atténuer** *Abaisser le prix des denrées* ⇒ **Diminuer**. *Abaisser la voix. — Abaisser un seuil, un temps Abaisser l'âge de la retraite* : réduire le nombre d'années de travail nécessaires, dans une profession donnée, au droit à la retraite ⇒ Abaissement, cit. 1.2
> Alg *Abaisser le degré d'un polynôme, d'une équation*, le ou la réduire à un degré inférieur.

La partie analyse de cette tâche a été effectuée et sa mise en œuvre est en cours.

Conclusion

Les cas de sous-catégorisation et les restrictions de sélection sont deux informations qui ne sont que très rarement notées dans les dictionnaires usuels et qui permettent de résoudre le problème du rattachement des groupes prépositionnels dans un analyseur.

Nous avons montré comment les cas de sous-catégorisation peuvent être automatiquement extraits de l'analyse des exemples de dictionnaires sur support informatique ou de l'analyse de corpus. Pendant la période durant laquelle nous avons effectué cette recherche, notre méthodologie a évolué : d'un système écrit sur mesure pour analyser les exemples du ZYZOMYS, nous en sommes venus à utiliser certains des outils linguistiques qui sont maintenant disponibles. L'utilisation de ces outils rend le travail moins fastidieux et permet de traiter un plus grand volume de données. Dans les deux cas cependant les données présentées au programme doivent être filtrées au préalable à défaut d'avoir un informateur disponible en ligne pour résoudre l'ambiguïté du rattachement. Fondé sur les travaux de Tomita (1984), nous avons construit la maquette d'un programme qui pose des questions à l'utilisateur dans le but de lever l'ambiguïté (Bouchard & Emirkanian, 1990).

Nous avons présenté un modèle de désambiguïsation des catégories lexicales en contexte qui se fonde sur un modèle discret des n-grammes qui nous semble être un remplaçant possible des algorithmes basés sur les modèles de Markov cachés, du moins pour les corpus qui nous intéressent. Partant d'un texte catégorisé et désambiguïsé en contexte, nous avons présenté un algorithme de segmentation simple et robuste qui permet de mettre en relief les arguments du verbe.

Finalement, nous avons illustré la façon de rendre compte des restrictions de sélection à partir d'une ontologie et de l'opération d'unification des types dans le treillis de l'ontologie.

Génération de dictionnaires-machines multilingues pour la traduction automatique de diagnostics médicaux

Guy DEVILLE et Emmanuel HERBIGNIAUX

École de langues vivantes, Facultés universitaires de Namur, Belgique

1. Introduction

Le projet ANTHEM[1] est le fruit de la collaboration entre chercheurs et industriels issus de diverses disciplines : médecine, informatique et linguistique[2]. Son objectif est de développer un prototype portable d'interface en langage naturel permettant la traduction et l'encodage automatiques de diagnostics médicaux (Ceusters *et al.*, 1994).

Cet article expose l'approche adoptée dans le développement des ressources lexicales multilingues du prototype ANTHEM, approche qui a nécessité une analyse minutieuse de grands corpus multilingues représentatifs de diagnostics médicaux.

Dans la section 2, nous décrivons les travaux de mise en forme et d'échantillonnage effectués sur les corpus de diagnostics médicaux. Ensuite, nous présentons dans la section 3 le modèle de représentation sémantique élaboré sur base des observations effectuées sur les échantillons. La section 4 est consacrée à une présentation sommaire de l'architecture du prototype ANTHEM. Nous abordons dans la section 5 le développement, l'extension et la maintenance proprement dits des lexiques électroniques parallèles français et néerlandais.

En guise de conclusion, nous évoquerons dans la section 6 les perspectives futures du projet ANTHEM, telles que le raffinement du modèle, la validation interne et

1. Le projet ANTHEM « Advanced Natural Language Interface for Multilingual Text Generation in Healthcare » (LRE 62-007) est co-financé par l'Union Européenne dans le cadre du Programme LRE (Linguistic Research and Engineering).

2 Le consortium ANTHEM est coordonné par RAMIT vzw/Gent (W. Ceusters) et comprend en outre les partenaires suivants FUNDP/Namur (G. Deville), CRP-CU/Luxembourg (P. Mousel), IAI/Saarbrücken (O. Streiter), ULG/Liège (C. Gérardy), Datasoft Management NV/Oostende (J. Devlies) et l'Hôpital militaire/Bruxelles (D Penson)

sur site du prototype ou encore les possibilités d'extension vers d'autres projets de recherches connexes ou d'autres applications.

2. Corpora de diagnostics médicaux

Toute entreprise de développement d'interfaces pour le traitement du langage naturel dans le domaine des soins de santé nécessite la modélisation du langage médical. Une telle tâche de modélisation exige l'observation d'un ensemble représentatif d'expressions du sous-langage de l'application. C'est donc dans un souci de validation empirique que deux corpus de diagnostics médicaux ont été collectés dans deux types d'environnements cliniques.

Un premier corpus est constitué de diagnostics établis et enregistrés sous format électronique par l'armée belge (corpus ABL). Ce premier corpus contient un total de 227 000 diagnostics en français et en néerlandais, rédigés par des médecins militaires entre 1970 et 1993, durant leurs consultations de militaires de carrière et de civils effectuant leur service militaire.

Parallèlement au corpus ABL, un corpus de 12 671 expressions françaises et néerlandaises a également été constitué (corpus Médidoc). Ce corpus couvre approximativement la même période que le corpus ABL. Il a été réalisé en réutilisant certaines parties de dossiers médicaux électroniques produits par un programme de gestion de données médicales (Médidoc) développé par la société Datasoft Management. Ce programme est utilisé par des médecins civils dans le cadre de leurs consultations privées.

Nous disposons donc au total de 239 671 diagnostics médicaux. Cependant, dans une perspective d'analyse et de modélisation linguistique, il était impensable d'observer les deux corpus dans leur ensemble. C'est pourquoi un premier échantillon de 2 343 expressions a été créé sur base des corpus ABL et Médidoc, dans lesquels chaque diagnostic avait, au préalable, été étiqueté à l'aide d'information concernant son origine, la langue et l'année de rédaction.

Ensuite, la validité linguistique et médicale de ces expressions a été vérifiée respectivement par des linguistes et des médecins. Il en a résulté un échantillon final de 1 362 expressions valides, utilisé pour *i*) l'élaboration de lexiques et de grammaires électroniques ainsi que pour *ii*) le testing du prototype ANTHEM au fil de sa réalisation. On trouvera dans Deville et Herbigniaux (1994) une description détaillée des principes méthodologiques appliqués dans l'élaboration des corpus de diagnostics médicaux du projet ANTHEM. C'est notamment sur base de cet échantillon final que le modèle de représentation sémantique des diagnostics médicaux a été élaboré et validé. Nous décrivons brièvement ce modèle dans la section suivante.

3. Modèle de représentation sémantique

Dans la tradition des grammaires casuelles, et plus spécialement de la grammaire fonctionnelle de Dik (1989), nous appelons *formule* la représentation sémantique d'un diagnostic. Une formule est une structure qui consiste en un *prédicat* et un nombre déterminé de *termes*, arguments du prédicat. Un prédicat est une expression (le plus

souvent un groupe nominal, adjectif ou verbe) reflétant les propriétés sémantiques de ses arguments ainsi que les relations sémantiques entre ces arguments. Un terme est une expression (le plus souvent un groupe nominal ou prépositionnel) faisant référence à une entité ou un ensemble d'entités dans un univers conceptuel de sous-langage (un sous-langage étant défini ici comme un ensemble d'expressions faisant référence à un domaine conceptuel limité et défini, et utilisé dans une fonction spécifique). Par conséquent, une formule fait référence à une *configuration d'univers de sous-langage*. Une configuration d'univers de sous-langage est une constellation d'entités conceptuelles du sous-langage exprimées en termes de leurs relations mutuelles (Deville, 1989).

Dans le cadre d'ANTHEM, les entités conceptuelles du sous-langage étudié sont définies comme étant les unités minimales qui ne font pas seulement référence à des éléments atomiques (*bras*) et complexes (*paume main gauche*), mais également à des états (*inflammation*), des processus (*tomber*) et des actes (*ingestion*). Plus précisément, les termes du langage d'application d'ANTHEM font référence à des objets, et les prédicats à des états, relations, actions et procès.

3.1. Typologie de prédicats

Les prédicats sont sélectionnés parmi une liste finie de *types sémantiques*. Un type sémantique capture les propriétés sémantiques et combinatoires prototypiques qui sont partagées par un ensemble de prédicats. La plupart des types sémantiques repris dans le modèle de représentation d'ANTHEM sont hérités d'une nomenclature standard largement répandue dans le domaine de la terminologie médicale : SNOMED (*Systematized Nomenclature of Medicine*) (CAP, 1993). Cette terminologie est reprise dans le tableau 1 ci-dessous.

disdia	disease/diagnosis
morpho	morphology
funct	function
topo	topography
modifiers	modifiers & general linkage
modalities	cfr. modifiers & general linkage
livor	living organisms
chemic	chemicals, drugs ...
physic	physical agents ...
proced	procedures
soct	social context
occup	occupations
hum	human (type ajouté)
temp	temporal (type ajouté)

TABLEAU 1 : Types sémantiques du modèle ANTHEM, selon la nomenclature SNOMED International.

3.2. Système casuel

Dans une formule, la relation entre le prédicat et ses arguments est spécifiée au moyen d'un *cas*. Un cas est l'expression d'une fonction ou d'un rôle sémantique prototypique rempli par le terme d'un prédicat dans la classe sémantique dont est issu ce prédicat. La *structure casuelle* d'un type sémantique est la séquence de cas nécessaires à la définition d'un ensemble de configurations d'univers de sous-langage, tel que représenté par ce type sémantique. Ainsi, un prédicat et ses arguments font référence à une configuration d'univers de sous-langage particulier, comme illustré dans l'exemple ci-dessous, qui reprend la représentation sémantique de l'expression *rupture du ménisque interne genou droit* :

```
(Rupture [Meniscus {medial} [Knee {right}]])
```

Un type sémantique associé à sa structure casuelle fait référence, à un niveau conceptuel plus prototypique, à une classe de configurations d'univers de sous-langage, comme le montre l'exemple ci-dessous. Une telle structure de 'haut niveau' spécifie les rôles sémantiques des arguments du prédicat, en relation avec le type sémantique correspondant à ce dernier :

```
(MORPHO -loc-> [TOPO {POSIT} -loc-> [TOPO {BODSID}]])
```

Une formule peut être étendue au moyen d'arguments périphériques. Les arguments périphériques ne participent pas en tant que tels à la définition d'une configuration d'univers de sous-langage, mais en expriment les dimensions spatio-temporelles, spécifient les entités secondaires qui participent à la configuration, ou encore précisent la manière, les conditions dans lesquelles cette configuration a lieu. Contrairement aux arguments centraux, les fonctions sémantiques des arguments périphériques ne sont pas nécessaires à la définition d'un ensemble de configurations en termes d'un type sémantique et de sa structure casuelle associée. La figure 1 illustre la représentation sémantique de l'expression *rupture du ménisc· ·e interne genou droit*, selon le modèle décrit plus haut.

Pour comprendre comment *i*) le modèle décrit plus haut et *ii*) les lexiques parallèles français et néerlandais sont implémentés dans le prototype ANTHEM, il est nécessaire d'en décrire brièvement l'architecture dans la section 4.

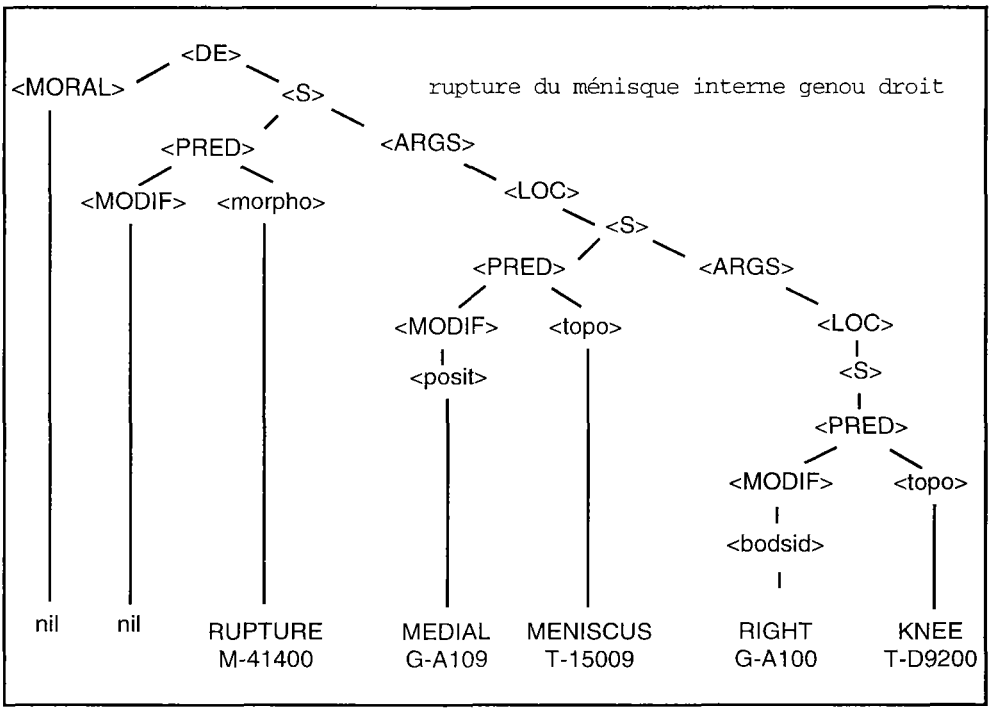

FIGURE 1: Représentation sémantique de l'expression
RUPTURE DU MENISQUE INTERNE GENOU DROIT.

4. Architecture du prototype ANTHEM

Le prototype ANTHEM est conçu pour pouvoir être intégré dans tout type d'application médicale existante, appelée application hôte. Rappelons que la fonction de ce prototype est de traduire des diagnostics médicaux exprimés en langage naturel, soit en une langue naturelle cible, un code ICD-10 ou une représentation sémantique. Pour les besoins de traduction automatique, le prototype fait appel au système de traduction CAT2, développé par l'IAI de l'Université de Saarbrücken, en marge du projet EUROTRA (Streiter *et al.*, 1994). En outre, l'encodage automatique de diagnostics vers le code ICD-10 a nécessité l'élaboration d'un système expert par le Laboratoire de Recherche en Informatique et Télématique Médicale de l'Hôpital Universitaire de Gand. Pour assurer l'interaction entre l'application hôte d'une part, et le système CAT2 et le système expert d'autre part, une interface (appelée *Application Programming Interface*) a été réalisé par le Centre de Recherche Public du Centre Universitaire de Luxembourg (Mousel et Thienpont, 1994). La figure 2 illustre l'architecture générale du prototype ANTHEM :

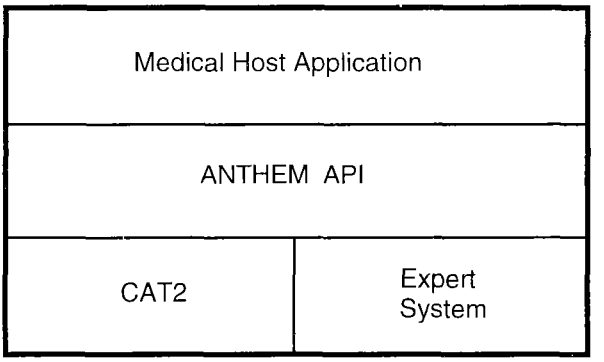

FIGURE 2 · Architecture générale du prototype ANTHEM

Avant d'aborder l'implémentation des lexiques et grammaires électroniques, nous exposons les principes de fonctionnement de CAT2. CAT2 est un système de traduction 'généraliste' basé sur l'unification de traits, et écrit en Prolog. Il nécessite *i*) la réalisation de modules de grammaire propres à chaque langue ainsi que *ii*) le développement de lexiques électroniques, également spécifiques à chaque langue. Les règles de grammaires sont en fait appliquées aux entrées conceptuelles présentes dans les lexiques et plus précisément à celles dont il est fait usage dans le diagnostic médical entré par l'utilisateur. Ces règles sont principalement de trois ordres, à savoir morphologique, syntaxique et sémantique.

Dans le cadre du projet ANTHEM, CAT2 a été adapté en vue de *i*) la traduction de diagnostics en langage naturel, ou *ii*) la transformation de ces expressions en une représentation sémantique indépendante de toute langue. C'est pour répondre à ces objectifs que des lexiques et grammaires électroniques spécifiques ont été implémentés. Au stade actuel du projet, des grammaires et des lexiques ont été produits pour le français, le néerlandais et, dans une moindre mesure, pour l'allemand. Ajoutons que l'essentiel du modèle de représentation sémantique du prototype ANTHEM est implémenté dans un module de grammaire commun à toutes les langues. Nous exposerons en détail dans la section 5 l'élaboration des lexiques électroniques dans ANTHEM.

5. Développement, extension et maintenance de lexiques électroniques

Dans cette partie, nous aborderons *i*) le développement de lexiques basé sur les corpus décrits plus haut, *ii*) les effets de l'intégration d'une composante morphologique plus élaborée dans le prototype ANTHEM sur l'extension et la maintenance des lexiques, *iii*) la typologie d'entrées lexicales actuellement en vigueur dans le projet ANTHEM et *iv*) l'extension semi-automatique du lexique, sur base de la version électronique du système de codification SNOMED International.

5.1. Développement de lexiques basé sur les corpus

Une des caractéristiques fondamentales des lexiques développés dans le cadre du projet ANTHEM est leur organisation que l'on peut qualifier de conceptuelle. En effet,

des termes comme *poumon* et *pulmonaire* sont considérés comme les réalisations substantivale et adjectivale d'un seul et même concept, celui de POUMON. Une telle approche conceptuelle des lexiques ANTHEM résulte en fait de l'organisation – également conceptuelle – du système international de codification SNOMED, utilisé comme interlangue dans notre projet.

Initialement, les lexiques français et néerlandais ont été développés sur base de quatre échantillons-tests, contenant chacun 50 expressions. Au cours de cette première phase, 215 concepts ont été implémentés dans chaque langue. Ensuite, toujours sur base d'expressions issues des corpus, le nombre d'entrées lexicales/conceptuelles a été porté à 586, dans le but de couvrir l'intégralité des 1 362 expressions sélectionnées au hasard et considérées comme valides, tant d'un point de vue linguistique que médical (échantillon final).

Notons toutefois que le lexique allemand a été élaboré et étendu sur base des entrées déjà présentes dans les lexiques français et néerlandais.

5.2. Intégration d'un composant morphologique plus élaboré

Au cours du projet, il a été décidé d'intégrer au prototype ANTHEM un module d'analyse et de génération morphologiques plus performant, dans le but de résoudre trois difficultés, à savoir *i*) l'analyse et la génération automatiques de formes plurielles (substantifs) et fléchies (adjectifs) (morphosyntaxe), *ii*) l'intégration d'informations utiles au traitement de la composition, de la négation et de la gradation (adjectifs) de termes en néerlandais (morphologie productive) et enfin *iii*) la reconnaissance automatique de variantes orthographiques en néerlandais (uniquement en analyse).

L'intégration d'un tel module nous a amené à optimiser le développement des lexiques ANTHEM. Initialement, l'information linguistiquement et médicalement relevante pour la constitution de ces lexiques était enregistrée sous forme de tables statiques. La génération des lexiques consistait alors à remplir manuellement des structures vides écrites en Prolog, à l'aide de l'information contenue dans ces tables. Outre son caractère répétitif, le principal inconvénient d'une telle approche est l'absence de lien actif et systématique entre les tables et les lexiques, la mise à jour de ces derniers nécessitant une double modification (tables et lexiques). D'autre part, toute modification affectant le format proprement-dit des entrées lexicales/conceptuelles (structures Prolog) doit être répétée quasi manuellement partout où cela s'avère nécessaire.

Afin de remédier au décalage sans cesse croissant entre les tables et les lexiques, dûs aux fréquentes modifications du lexiques sans modifications équivalentes dans les tables, nous avons procédé à l'extraction automatique de l'information relevante des lexiques, en vue de constituer une version tout à fait à jour des bases de données. Notons que cette procédure ne remédiait pas encore aux inconvénients exposés plus haut.

C'est alors que nous avons opté pour une approche en sens inverse, à savoir de constituer les lexiques de manière automatique sur base d'information linguistiquement et médicalement pertinente contenue dans des bases de données évolutives, dans la mesure où les modifications touchant aux entrées conceptuelles proprement-dites sont uniquement effectuées dans ces bases. À cette fin, une série de générateurs de

lexique ont été développés en langage de programmation AWK (Aho *et al.*, 1988), afin de pouvoir, à partir de ces bases de données, créer et adapter automatiquement des entrées conceptuelles de plusieurs types, tant en français qu'en néerlandais.

Les avantages d'une telle approche sont au nombre de quatre : tout d'abord, le développeur de tels lexiques électroniques peut se concentrer uniquement sur l'acquisition et la maintenance d'informations linguistiquement et médicalement relevantes ; ensuite, on obtient un haut degré d'adaptabilité du format d'implémentation ainsi qu'une portabilité éventuelle de l'information pertinente vers d'autres formalismes ; on peut également générer de manière quasi automatique de grandes quantités d'entrées lexicales/conceptuelles ; enfin, on peut mettre au point une série de procédures de vérification automatique de cohérence de l'information. Il est en outre possible d'effectuer rapidement divers relevés statistiques en observant les données selon différents axes. Notons que ces procédures de vérification et ces programmes de relevés statistiques sont également écrits en AWK. Dans la section suivante, nous examinons en détail l'organisation interne des lexiques électroniques d'ANTHEM.

5.3. Typologie d'entrées lexicales

Actuellement, quatre types d'entrées sont utilisés dans les lexiques ANTHEM, à savoir *i*) les termes simples, *ii*) les nombres – cardinaux et ordinaux –, *iii*) les pseudo-composés et *iv*) les *multi word units* (concepts constitués de plusieurs mots séparés par des blancs). À cette fin, quatre générateurs de lexique ont été développés pour chaque langue, en vue de générer automatiquement des entrées *i*) simples, *ii*) numérales, *iii*) pseudo-composées et *iv*) à mots multiples.

Notons encore que, parmi les entrées simples, on distingue des concepts réalisés syntaxiquement au moyen *i*) d'un ou plusieurs substantif(s), *ii*) d'un ou plusieurs adjectif(s) ou encore *iii*) d'un ou plusieurs substantif(s) et d'un ou plusieurs adjectif(s).

Parmi les entrées à mots multiples, on distingue actuellement *i*) les entrées constituées de deux mots (le plus souvent un substantif et un adjectif) et *ii*) celles constituées de trois mots (le plus souvent un substantif, une préposition et un nom propre). Nous présentons brièvement ci-dessous quelques exemples illustrant les principaux types d'entrées lexicales en français.

5.3.1. Entrées simples

Dans les exemples ci-dessous, *after_effect*, *chronic* et *fracture* sont utilisés comme aide-mémoire par le développeur. Ces étiquettes ne sont pas traitées par le système CAT2. Les valeurs *F-01460*, *G-A270* et *M-12000* (variable *lex*=) sont les codes SNOMED correspondant aux concepts de SÉQUELLE, CHRONIQUE et FRACTURE. Les valeurs *funct*, *progr* et *morpho* (variable *type*=) indiquent la catégorie sémantique – inspirée de la typologie SNOMED – à laquelle les différents concepts appartiennent. Outre la réalisation syntaxique des différents concepts en français (variable *string*=), nous trouvons aussi de l'information concernant la catégorie syntaxique (variable *cat*=), le genre (variable *gen*=) et le profile morphologique des différents mots implémentés (variable *flex*=).

```
after_effect =
{lex='F-01460',type=funct,
 known=yes,
 head={cat=n,ehead={cat=n}}}>>
({string=se1quelle,
 onset=cons,
 flex=normal,
 head={cat=n,ehead={gen=fem}}}).[].

chronic =
{lex='G-A270',type=progr,
 known=yes,
 head={cat=a,ehead={cat=a}}}>>
({string=chronique,
 onset=cons,
 flex=normal,
 head={cat=a,apos=post}}).[].

fracture =
{lex='M-12000',type=morpho,
 known=yes,
 head=({cat=n,ehead={cat=n}}
      ;{cat=a,ehead={cat=a}})}>>
({string=fracture,
  onset=cons,
  flex=normal,
  head={cat=n,ehead={gen=fem}}}
;{string=fract,
  anasyn=ana,
  onset=cons,
  flex=normal,
  head={cat=n,ehead={gen=fem}}}
;{string=casse1,
  onset=cons,
  flex=normal,
  head={cat=a,apos=post}}
;{string=fracture1,
  anasyn=ana,
  onset=cons,
  flex=normal,
  head={cat=a,apos=post}}).[].
```

5.3.2. Entrées numérales

Les principes exposés ci-dessus restent d'application pour les entrées de type numérales, bien que le code SNOMED soit remplacé ici par un code plus transparent (variable *lex=*). Il est à noter que les réalisations cardinales et ordinales renvoient au même concept.

```
ruleEIGHTEEN =
 {lex='18',type=quant,clex=nil,
  known=yes,
  head={cat=a,ehead={cat=a}}}>>
```

```
({string='dix-huit',
 onset=cons,
 flex=inv,
 head={cat=a,apos=pre}}
;{string=xviii,
 anasyn=ana,
 onset=cons,
 flex=inv,
 head={cat=a,apos=pre}}).[].

ruleEIGHTEENTH =
 {lex='18',type=ind,clex=nil,
 known=yes,
 head={cat=a,ehead={cat=a}}}>>
({string='dix-huitie2me',
 onset=cons,
 flex=inv,
 head={cat=a,apos=pre}}
;{string='18e2me',
 anasyn=ana,
 onset=cons,
 flex=inv,
 head={cat=a,apos=pre}}
;{string='18e',
 anasyn=ana,
 onset=cons,
 flex=inv,
 head={cat=a,apos=pre}}
;{string=xviiie2me,
 anasyn=ana,
 onset=cons,
 flex=inv,
 head={cat=a,apos=pre}}
;{string=xviiie,
 anasyn=ana,
 onset=cons,
 flex=inv,
 head={cat=a,apos=pre}}
;{string='dix-huit',
 anasyn=ana,
 onset=cons,
 flex=inv,
 head={cat=a,apos=post}}
;{string=xviii,
 anasyn=ana,
 onset=cons,
 flex=inv,
 head={cat=a,apos=post}}).[].
```

5.3.3. Entrées pseudo-composées

Les entrées de type pseudo-composé, ainsi que celles à mots multiples présentées plus bas, sont certainement les plus intéressantes. Le principe des entrées pseudo-composées

est de rendre explicite la structure conceptuelle interne d'un mot donné. On ne procède cependant à une telle explicitation que si celle-ci est motivée tant d'un point de vue médical que linguistique. En effet, la majorité des mots d'un sous-langage donné peuvent se décomposer en une structure conceptuelle complexe. Dans le sous-langage des diagnostics médicaux, par exemple, on trouve de nombreux affixes qui ont une signification particulière à l'intérieur des mots dans lesquels ils apparaissent. Par exemple, le suffixe *-ite* (en français) ou *-itis* (en néerlandais) signifie très souvent *inflammation localisée dans/affectant 'xxx'*, 'xxx' étant exprimé dans le(s) premier(s) morphème(s) du mot.

Mais, dans le projet ANTHEM, seuls les termes nécessitant une telle explicitation de leur structure conceptuelle interne sont implémentés sous forme de pseudo-composés. Nous appelons ces termes pseudo-composés afin de les distinguer d'un autre type de composés, à savoir les composés tels que *laryngotrachéite*, dans lesquels on peut aisément reconnaître *laryngite* (ou *larynx*) et *trachéite*, et qui, dès lors, peuvent être traités automatiquement par l'analyseur morphologique du système CAT2.

Le terme *trigéminie* (inflammation du nerf facial appelé trijumeau) illustre bien le principe des entrées de type pseudo-composé. Dans cet exemple, il a fallu rendre explicite dans le lexique la structure conceptuelle de ce terme car on rencontre dans le sous-langage médical des expressions telles que *trigéminie gauche*. Or, pour interpréter correctement le terme *gauche* dans cette expression, il faut savoir que l'on réfère en fait au *nerf trijumeau gauche* (par opposition au droit), concept non réalisé grammaticalement ici, suite à un phénomène d'ellipse 'conceptuelle'. Le modificateur *gauche* ne qualifie donc pas le terme *trigéminie* au même titre que les adjectifs *aigüe*, *récidivante* ou *bénigne*, par exemple. L'entrée lexicale du concept TRIGÉMINIE aura donc la forme suivante :

```
inflammation_in_trigeminalnerve =
{lex='M-40000',type=morpho,
clex={lex='T-A8150',type=topo,
      role=loc,head={ehead={cat=n,pform=in,feature=in}}},
known=yes,
head={cat=n,ehead={cat=n}}>>
({string=trigeminie,
  onset=cons,
  flex=normal,
  head={cat=n,ehead={gen=fem}}}).[].
```

Dans l'exemple ci-dessus, le code *M-40000* identifie le concept d'INFLAMMATION, tandis que le code *T-A8150* renvoie au concept de NERF TRIJUMEAU. Le lien entre les deux est exprimé à l'aide de la valeur *loc* (variable *role=*) ainsi que des paramètres *pform=in* et *feature=in* dont nous n'expliquerons pas le fonctionnement ici.

D'autres exemples d'entrées pseudo-composées sont : *gonarthrose* (arthrose localisée dans le genou), *tenniselbow* (inflammation localisée dans le coude), *épicondylite* (inflammation localisée dans l'épicondyle) ou encore *gonalgie* (inflammation localisée dans le genou).

5.3.4. Entrées à mots multiples

Dans cette sous-section, on trouvera, à titre d'exemple, l'implémentation des entrées à mots multiples référant aux concepts de CRÊTE ILIAQUE (substantif + adjectif) ainsi que de MALADIE DE SCHEUERMANN (substantif + préposition + nom propre).

```
crista_iliaca =
 {lex='T-1234A',type=topo,
  known=yes,
  head={cat=n,ehead={cat=n,gen=GEN,num=NUM}},
  mwu1={lex=iliaca,head={ehead={gen=GEN,num=NUM}}}}>>
({string=cre3te,
  onset=cons,
  flex=normal,
  head={cat=n,ehead={gen=fem}}}).[].

iliaca =
 {lex=iliaca,type=mwu,role=mwu,
  known=yes,
  head={cat=a,ehead={cat=a}}}>>
({string=iliaque,
  onset=vowel,
  flex=normal,
  head={cat=a,apos=post}}).[].
```

Dans l'exemple ci-dessus (*crête iliaque*), le terme principal est *crête* (en fait 'cre3te', les accents étant implémentés sous la forme de combinaisons voyelle + chiffre). Le concept CRÊTE ILIAQUE, de type TOPO (variable *type=*), est identifié par le code SNOMED T-1234A (variable *lex=*).

Le lien entre le substantif 'crête' (*cat=s*) et l'adjectif 'iliaque' (*cat=a*) est ensuite établi à l'aide de la variable *mwu1=*, comportant entre autres, dans sa structure sous-jacente, le trait *lex=iliaca*, aide-mémoire utilisé pour désigner l'implémentation du terme 'iliaque', de type MWU (variable *type=*).

L'accord en genre et en nombre entre le substantif (*crête*) et l'adjectif (*iliaque*) est garanti par les traits *gen=GEN* et *num=NUM*, dans lesquels *GEN* et *NUM* sont des variables dont les valeurs doivent être identiques pour *crête* (terme principal) et pour *iliaque* (mwu).

D'autres exemples d'entrées à mots multiples constituées d'un substantif et d'un adjectif sont : *ongle incarné, nodule froid, corps étranger, nœud atrioventriculaire, veine saphène, tronc basilaire, globule rouge, globule blanc* ou encore *cage thoracique.*

L'implémentation du concept de MALADIE DE SCHEUERMANN illustre un second type d'entrée à mots multiples :

```
scheuermann_disease =
 {lex='D1-61210',type=disdia,
  known=yes,
  head={cat=n,ehead={cat=n}},
  mwu1={lex=de},
  mwu2={lex=scheuermann}}>>
({string=maladie,
  onset=cons,
  flex=normal,
  head={cat=n,ehead={gen=fem}}}
;{string=syndrome,
  anasyn=ana,
  onset=cons,
  flex=normal,
  head={cat=n,ehead={gen=masc}}}).[].

de =
 {lex=de,type=mwu,role=mwu,
  known=yes,
  head={cat=mwu,ehead={cat=mwu}}}>>
({string=de,
  head={cat=mwu}}).[].

scheuermann =
 {lex=scheuermann,type=mwu,role=mwu,
  known=yes,
  head={cat=mwu,ehead={cat=mwu}}}>>
({string='Scheuermann',
  onset=cons,
  head={cat=mwu}}).[].
```

Dans l'exemple ci-dessus, le terme principal est *maladie/syndrome*. Le concept de MALADIE DE SCHEUERMANN, de type DISDIA (variable *type=*), est identifié par le code SNOMED D1-61210 (variable *lex=*).

Le lien entre le substantif 'maladie'/'syndrome' (*cat=s*) et les autres éléments, à savoir la préposition 'de' et le nom propre 'Scheuermann' (ayant tous deux les traits *type=mwu* et *cat=mwu*) est ensuite établi à l'aide des variables *mwu1* et *mwu2*, comportant respectivement les traits *lex=de* et *lex=scheuermann*, aides-mémoires utilisés pour désigner l'implémentation des termes 'de' et 'Scheuermann'.

Notons enfin qu'aucune variation morpho-syntaxique n'est prévue, ni pour *de*, ni pour *Scheuermann*. En outre, l'implémentation de la préposition *de* est utilisée dans plusieurs entrées à mots multiples, comme par exemple les *maladies* et/ou *syndromes de Mallory-Weiss, Becker* ou *Crohn*, la *fracture de Pouteau-Colles* ou encore la *maladie de système*, l'*accident de voiture*, la *tête de radius*, l'*accès de panique*, le *bouchon de cérumen*, l'*eczéma de contact*, la *fracture de stress*, le *mal de tête*, l'*hydrate de carbone* ou l'*acétate de désoxycorticostérone*.

5.4. Extension semi-automatique du lexique

Dans un premier temps, les lexiques français et néerlandais ont été développés dans le

but de pouvoir analyser et générer un nombre croissant d'expressions issues d'échantillons représentatifs des corpus ABL et Médidoc. C'est ainsi qu'un total de 586 entrées simples avait été implémenté.

Dans la phase d'extension intensive des lexiques ANTHEM, nous avons opté pour une certaine systématisation dans la création de nouvelles entrées. Nous avons en effet décidé d'implémenter le plus possible d'entrées simples ou atomaires appartenant essentiellement aux cinq catégories de la typologie SNOMED les plus fréquemment utilisées, à savoir DISDIA (maladies/affections), MORPHO (morphologie), TOPO (topographies/endroits du corps), GLMOD (modificateurs) et FUNCT (fonctions).

Outre le développement systématique des axes prototypiques de la nomenclature SNOMED International, le principal avantage d'une telle approche par rapport à une démarche inspirée de l'observation des corpus est évident : pour tout mot nouveau apparaissant dans un ordre aléatoire, il n'est plus nécessaire de rechercher le code correspondant au concept que ce mot réalise, ni la catégorie sémantique à laquelle ce concept appartient. Ces informations sont en effet présentes de manière systématique dans la version électronique de SNOMED dont nous sommes partis pour l'extension semi-automatique du lexique.

La phase d'extension intensive des lexiques a permis l'implémentation de 467 entrées simples appartenant à la catégorie DISDIA, 550 à la catégorie MORPHO, 342 à la catégorie TOPO, 376 à la catégorie GLMOD et 1 146 à la catégorie FUNCT, soit un total de 3 590 entrées, auxquelles se sont ajoutées 26 entrées pseudo-composées et 115 entrées à mots multiples, tant en français qu'en néerlandais.

6. Conclusion et perspectives futures

Dans cet article, nous avons exposé l'approche adoptée dans le développement des ressources lexicales multilingues du prototype ANTHEM, approche qui a nécessité une analyse minutieuse de grands corpus multilingues représentatifs de diagnostics médicaux.

Nous avons décrit les travaux de mise en forme et d'échantillonnage effectués sur les corpus de diagnostics médicaux. Nous avons également présenté le modèle de représentation sémantique élaboré sur base des observations effectuées sur les échantillons et exposé l'architecture du prototype ANTHEM. Enfin, nous avons abordé le développement, l'extension et la maintenance proprement-dits des lexiques électroniques parallèles français et néerlandais dans les différentes phases de leur élaboration, dans une perspective d'automatisation de la plupart de ces tâches.

Au travers du projet ANTHEM, nous avons tenté d'identifier les fondements d'une modélisation adéquate du langage naturel dans le domaine des soins de santé, en mettant l'accent sur la nécessité d'une validation à la fois empirique, linguistique et opératoire des modèles obtenus. Une telle approche nous semble répondre aux exigences de la multidisciplinarité évoquée plus haut : en effet, la modélisation lexicale du langage d'application d'ANTHEM *i*) est élaborée à partir d'expressions générées

dans un environnement clinique en vraie grandeur, *ii*) met en œuvre les dimensions linguistiques caractérisant la plupart des 'États de Choses' grammaticalisés dans tout diagnostic médical, et *iii*) est exprimé à l'aide d'un formalisme permettant à toute expression en langage naturel d'être directement 'calculable' par un module de traduction automatique.

Pour conclure, nous évoquons brièvement quelques perspectives futures du projet ANTHEM. Tout d'abord, le modèle de représentation sémantique du langage de l'application sera sans doute encore raffiné. Les éléments nécessaires pour ce raffinement seront obtenus essentiellement grâce aux validations internes et sur site du prototype ANTHEM, qui sera alors confronté à de nouveaux diagnostics médicaux issus des corpus ABL et Médidoc, ainsi qu'au milieu médical en vraie grandeur. Enfin, les possibilités d'extension vers d'autres projets de recherche connexes ou d'autres applications sont dès à présent envisagées avec le plus grand intérêt.