BUTS ET MÉTHODES DE L'ÉLABORATION DES DICTIONNAIRES ÉLECTRONIQUES DU LADL

Blandine COURTOIS,
LADL, CNRS, Université Paris 7.

"L'élaboration d'un dictionnaire général de la langue exige un travail assidu, poursuivi durant de longues années. Il faut, pour s'y astreindre, une foi persévérante dans l'utilité de l'effort."

Paul Robert. Introduction au Grand Robert de la Langue Française.

Ces mots de Paul Robert, placés en tête de la première édition de son Grand Dictionnaire, peuvent également servir d'introduction aux dictionnaires électroniques. En effet, ceux-ci comme celui-là sont le résultat d'un important travail de rassemblement, d'analyse, de présentation et de contrôles de données linguistiques en très grand nombre. Toutefois, si l'effort et la persévérance sont identiques pour les deux types de dictionnaires et si la description de la langue est leur objet commun, nous verrons que là s'arrête la comparaison et qu'on ne saurait utiliser un dictionnaire usuel mis sur support magnétique comme dictionnaire électronique.

1. PRÉSENTATION GÉNÉRALE DU SYSTÈME DELA

1.1. BUTS

Le système DELA des Dictionnaires Électroniques du LADL est un système linguistique et informatique, qui regroupe
- des modules de description de la langue française,
- des programmes de traitement des données.

Le but est l'analyse automatique des textes et, à plus long terme, la communication en langue naturelle avec l'ordinateur. Or ce dernier requiert des modèles formellement définis et une cohérence absolue dans les descriptions. C'est pour répondre à ces exigences que sont élaborés les dictionnaires électroniques, construits comme des bases structurées où les unités de la langue sont répertoriées sur leur aspect formel et où les propriétés linguistiques sont décrites sous une forme utilisable par l'ordinateur.

1.2. COMPARAISON AVEC LES DICTIONNAIRES USUELS

Face aux besoins de la machine, les dictionnaires traditionnels informatisés ne sont pas adaptés. D'une part les différents dictionnaires du commerce ne sont pas tous équivalents, ayant chacun une couverture lexicale limitée qui correspond aux objectifs visés par l'éditeur. D'autre part, tous supposent une compétence du lecteur, appelé à mettre en jeu implicitement ses connaissances du monde, sa faculté d'interprétation des définitions, sa compréhension des mots par analogie avec d'autres mots ou avec des mécanismes de création connus. Ainsi, les adjectifs en *-able*, dérivés de verbe (*délocaliser-délocalisable*), les préfixations en *re-* (*recoder*, *recodage*), les adverbes en *-ment*,... sont sous-représentés, même dans les grands dictionnaires. Du fait que la dérivation est régulière et intuitivement comprise, elle est omise. Sans doute, de telles entrées ne sont pas indispensables pour un lecteur compétent, mais il n'en va pas de même pour l'ordinateur. En

effet la première tâche de ce dernier, lors de l'analyse de textes, est la consultation des mots dans le dictionnaire. L'absence d'un seul mot entraîne alors l'échec de la consultation, aussi tous les mots valides rencontrés dans des textes et publications variées doivent-ils être représentés en entrée de dictionnaire électronique. D'où la nécessité d'une **large couverture lexicale,** tendant vers l'exhaustivité, bien que celle-ci ne puisse être réellement atteinte.

Si les entrées des dictionnaires du commerce ne peuvent suffire, le contenu des articles est également inexploitable pour un analyseur syntaxique. Certaines données comme la définition ou l'étymologie lui sont inutiles. D'autres comme les catégories de discours sont essentielles, mais ne sont pas toujours notées de façon identique dans toutes les entrées. Or, il est indispensable d'avoir une standardisation du format des données décrites, c'est-à-dire une **codification** systématique et rigoureuse des articles. Ceci signifie que les informations associées à chaque mot d'entrée doivent toutes être **explicites**, ou, si elles ne le sont pas, doivent pouvoir être calculées par un algorithme connu de l'ordinateur, explicitement formulé dans une partie du système.

En définitive, la différence entre dictionnaires traditionnels et dictionnaires électroniques pourrait être résumée ainsi :
- les dictionnaires d'usage, même transcrits sur support informatique, sont destinés à des lecteurs humains. Ils sont en conséquence orientés vers la définition des mots et la signification de leurs emplois.
- les dictionnaires électroniques sont construits pour l'ordinateur. Ils sont alors focalisés sur la **description formelle** des objets de la langue et leur **classification systématique**. Ce sont donc des ensembles très spécifiques de données qui sont ainsi élaborés et qui constituent le système DELA.

1.3. ORGANISATION DES TRAVAUX DU LADL.

Structurellement, le système DELA est organisé en plusieurs modules. La répartition des données dans chaque

module est effectuée d'après la forme des mots d'entrée et la nature des informations qui les accompagnent.

La configuration générale du système se compose des ensembles de données linguistiques suivants :
- dictionnaires de mots simples, DELAS et DELAF,
- dictionnaires de mots composés, DELAC et DELACF,
- dictionnaires phonémiques, DELAP et DELAPF,
- tables syntaxiques, regroupées en lexique-grammaire,
- graphes et automates.

A ces ensembles sont associés des outils de traitements informatiques. Entre autres :
- des programmes de génération de formes fléchies,
- des programmes de phonémisation automatique,
- un logiciel d'interrogation de textes, où sont intégrées de nombreuses fonctionnalités, et réalisé par Max Silberztein. Ce dernier, dans son ouvrage sur les dictionnaires électroniques, donne une description détaillée des possibilités offertes par ce logiciel, diffusé sous le nom INTEX.

L'exposé qui suit concerne uniquement la partie linguistique du système DELA. Tout d'abord, le contenu de chacune des structures de données linguistiques sera présenté succinctement. Puis quelques particularités et problèmes liés à l'élaboration des dictionnaires électroniques seront mis en relief.

2. STRUCTURES DE DONNÉES LINGUISTIQUES

2.1. DICTIONNAIRES DE MOTS SIMPLES

DELAS est le dictionnaire électronique des **mots formellement simples** du français. Ces mots sont mis sous leur forme canonique, dans l'ordre alphabétique, et sont suivis d'une codification systématique de leur catégorie grammaticale et d'un

code morphologique pour les mots variables. Chaque entrée est dotée en outre d'un numéro d'appartenance à une couche lexicale donnée. De plus, des marques sémantiques sont associées aux noms, et des renvois vers des tables de constructions syntaxiques accompagnent les verbes.

Exemples :

inviter,"1.V3(t;11)	Verbe de classe V3
précepteur,"1.N36(Hum)	Nom humain, classe N36
légal,"1.A76	Adjectif, classe A76.

Les codes morphologiques du DELAS renvoient à des classes flexionnelles, établies au préalable :

- conjugaisons (99 modèles),
- flexions nominales ou adjectivales (80 types),
- flexions rares (*tiers/tierce, oeil/yeux*,...)
- flexions avec alternatives (*lunch/ lunches+ lunchs*, *solo/ solos+ soli*,...)
- flexions de déterminants.

Chaque classe flexionnelle est décrite par une suite de terminaisons formelles et un numéro de code équivalent à cet ensemble de terminaisons. Ainsi, la classe de code A76 est équivalente à :

A76 = (*l,le,ux,les*), flexion qui permet de calculer par programme toutes les formes de *légal* et des mots de la même classe.

Actuellement les entrées de mots simples représentent plus de 90 000 graphies, toutes différentes. Une graphie sert d'entrée commune aux divers homographes existants. Par exemple, *déjeuner* a une seule entrée :

déjeuner,"1.N1.V3

qui rassemble un verbe et un nom se distinguant chacun par leur propre code. Les entrées à plusieurs codes concernent environ dix pour cent du nombre total de graphies, de sorte que le DELAS contient plus de 100 000 mots de codes grammatical ou morphologique différents.

DELAF est le dictionnaire électronique des **formes simples fléchies** du français. Chaque forme d'entrée est identifiée par la forme canonique et son code morphologique, et par des codes représentant :

- genre et nombre pour les noms, adjectifs, mots grammaticaux de forme variable,
- personne, genre, nombre pour les pronoms personnels,
- mode, temps, personne, nombre pour les formes de verbes.

Exemples :

préceptrice,*précepteur*. N36(Hum):fs
invitions,*inviter*. V3(t;11):IIM1p:SPR1p
légaux,*légal*.A76:mp
ils,.PRO(PpvIL):3mp

Le dictionnaire DELAF est construit automatiquement à partir du DELAS, par un programme de génération de formes fléchies. La procédure consiste à utiliser le code morphologique de chaque mot pour retrouver la classe flexionnelle correspondante, à partir de laquelle il devient possible d'engendrer toutes les formes fléchies.

Dans sa version la plus récente, le DELAF comporte de l'ordre de 750 000 formes simples fléchies, identifiées grammaticalement et par leur forme canonique.

2.2. DICTIONNAIRES DE MOTS COMPOSÉS

DELAC est le dictionnaire électronique des **mots composés** et de leur morphologie. Les entrées sont donc des unités lexicales formellement composées. Elles sont accompagnées de codes précisant leurs variations de formes, ainsi que d'indications de traits sémantiques.

Exemples :

cousin/germain,un/N32/A32/ms;++;Hum NA
pomme/de//terre,une/N21/fs;-+;Conc NDN
actualités/télévisées,les/fp;-- NA

L'ensemble du DELAC est subdivisé en sept classes, d'après les constituants caractéristiques de la composition :

- NA : nom/adjectif, tel *un cousin germain*,
- NDN : nom/de/nom, tel *une pomme de terre*,
- NAN : nom/à/nom, tel *un bateau à voile*,
- AN : adjectif/nom, tel *un petit-fils,*
- NN : nom/nom, tel *un homme-grenouille*,
- PN : préposition/nom, tel *un sans-gêne*,
- VN : verbe/nom, tel *un fume-cigare*.

Les mots composés de ces classes ont été collectés par différents linguistes travaillant en coopération avec le LADL, notamment Gaston Gross, Robert Vivès, Michel Mathieu-Colas, René Jung. Des adverbes composés, tels *par hasard, de longue date*, ont été aussi recensés par M. Gross et des conjonctions composées, telles *à cause de, en admettant que*, recueillies par M. Piot. Le regroupement des mots composés dans le DELAC, puis leur codification en genre, nombre et codes flexionnels, et enfin leur flexion, sont l'oeuvre de Max Silberztein.

Parmi les sept classes de composés mentionnées plus haut, les deux classes NA et NDN comptent le plus grand nombre de représentants. Au total 160 000 mots composés environ sont rassemblés et codés dans le DELAC.

DELACF est le dictionnaire électronique qui contient l'ensemble des **formes composées fléchies**. La flexion des mots composés est obtenue automatiquement à partir des codes flexionnels des unités simples qui permettent de calculer l'ensemble des variations de formes. Le fonctionnement de la flexion est dépendant de la classe du composé : par exemple dans la classe NA, les deux constituants sont généralement variables en nombre (*des cousins germains*), mais dans la classe NDN, le deuxième élément est le plus souvent invariable (*des pommes de terre*).

Ensemble les deux dictionnaires DELAF et DELACF recouvrent la totalité des formes françaises, simples et

composées. L'un et l'autre sont systématiquement consultés lors des procédures de reconnaissance des mots d'un texte.

2.3. DICTIONNAIRES PHONÉMIQUES

DELAP est un dictionnaire phonémique dont les entrées sont parallèles à celles du DELAS. Il comporte de plus, en regard de chaque mot d'entrée, une **représentation phonémique** de sa prononciation. Celle-ci correspond à un découpage des mots en syllabes abstraites en fonction des phonèmes qui les constituent.

Exemple :

discothèque,*/diskotek/*, N21

La zone phonémique, entre les deux virgules, donne donc la prononciation du mot sous forme d'une chaîne de phonèmes. Cette représentation est proche d'une transcription phonétique, mais plus informative. Construit par E. Laporte, le DELAP est largement développé dans sa thèse, en même temps que les méthodes algorithmiques et lexicales de phonétisation du français.

Le DELAP étant strictement parallèle au DELAS, le nombre de ses entrées est équivalent à celui du DELAS.

DELAPF est semblable à DELAP, mais pour les formes fléchies : il contient la représentation phonémique de la prononciation de l'ensemble des formes simples du français.

Les dictionnaires phonémiques offrent une description complète de la phonétique du vocabulaire courant. Outre leur intérêt pour les recherches linguistiques, ils ont déjà fait l'objet d'applications, notamment en correction orthographique par phonétisation. Leur apport est important pour la reconnaissance de la parole.

2.4. LEXIQUE-GRAMMAIRE

Le dictionnaire électronique contenant la **syntaxe** est désigné au LADL sous le terme **lexique-grammaire**. Il est

fondé sur la théorie transformationnelle de Z.S.Harris, selon laquelle l'unité de sens est la phrase, les transformations permettant de construire des classes d'équivalence par des procédures syntaxiques. Basé sur cette théorie, le lexique-grammaire est subdivisé en tables homogènes qui regroupent des éléments du lexique ayant un fonctionnement syntaxique comparable. La méthode consiste à définir des structures-types de construction pour chaque table, et à classer les éléments d'après ces structures. Ensuite, les propriétés syntaxiques, distributionnelles, sémantiques, sont codifiées en regard de chaque entrée de table.

Sur les méthodes en syntaxe, les ouvrages du professeur Maurice Gross sont des références essentielles, et nous y renvoyons le lecteur. L'élaboration de tables de constructions syntaxiques au LADL a bénéficié de nombreuses contributions de linguistes. La syntaxe des verbes français est entièrement décrite. Les adverbes figés et les conjonctions sont également bien étudiés. La syntaxe des noms est encore en phase d'élaboration.

Le lexique-grammaire des verbes comprend quatre types principaux de constructions de phrases : complétives, transitives, intransitives et locatives. Les tables syntaxiques, dues à Maurice Gross, Jean-Paul Boons, Alain Guillet et Christian Leclère, constituent une description complète des emplois des verbes français. Actuellement, la classification est répartie en 81 tables, contenant plus de 31 000 emplois, dont près de 20 000 phrases figées (recueillies par Maurice Gross). Des liens permettent de communiquer du DELAS vers ces tables.

2.5. GRAPHES ET AUTOMATES

Des représentations sous forme de graphes d'automates finis sont également intégrées au système DELA. Graphes et automates sont des représentations équivalentes. Sur le plan informatique, les algorithmes de traitements d'automates par l'ordinateur ont l'avantage d'être déjà largement étudiés. Sur le

plan linguistique, l'intérêt des graphes ou des automates est d'offrir un moyen de regrouper dans une même description des éléments équivalents du point de vue du sens, mais différents au niveau de la forme, par exemple les variantes graphiques de mots dont l'orthographe n'est pas normalisée (*bistrot* + *bistro*, *tsigane* + *tzigane*) ou des variantes d'expressions de temps (*avec le recul du temps* + *avec le recul des années*).

La reconnaissance de séquence de mots par automates a fait l'objet d'un travail de thèse de Denis Maurel. L'application aux adverbes de date du français a été testée avec succès, et le traitement des descriptions sous forme d'automates est inclus dans le logiciel INTEX.

2.6. PRODUITS DÉRIVÉS

Les données contenues dans les dictionnaires électroniques du LADL étant systématiquement codées, elles sont facilement exploitables pour obtenir des produits dérivés.

Notamment le DELAS est à l'origine de :
- dictionnaire **par parties du discours**,
  verbes : DELAS-V, noms : DELAS-N,
  adjectifs : DELAS-A adverbes : DELAS-ADV
- dictionnaire grammatical **en tri inverse**, DELAS-I, dans lequel les mots simples sont triés de droite à gauche. Le tri ainsi effectué regroupe entre eux les mots de même suffixe, souvent de même code morphologique
- dictionnaire **par classes flexionnelles**,
  verbes : DELAS-CV, noms : DELAS-CN,
  adjectifs :DELAS-CA

Des programmes spécifiques ont été construits pour analyser la structure des mots de la langue. D'autres ont permis de constituer des dictionnaires d'**anagrammes**, calculés à partir des mots simples, et à partir des mots composés.

Une recherche systématique de toutes les **formes ambiguës** du DELAF a également été menée, aboutissant à un répertoire

très complet d'ambiguïtés, important pour l'étude de mécanismes de levée automatique d'ambiguïtés en analyse de textes.

3. PARTICULARITÉS DU DICTIONNAIRE DELAS

Il n'est guère possible de développer plus ici les différents ensembles construits au LADL. La plupart d'entre eux ont donné lieu à des publications de leurs auteurs, et nous engageons le lecteur à consulter la bibliographie pour plus amples détails sur ces constructions. Néanmoins, afin de faire ressortir certains aspects particuliers du système DELA, nous utiliserons les dictionnaires DELAS et DELAF qui sont des modules de base du système. En effet, DELAF joue un rôle de dictionnaire de référence, en ce sens qu'il doit contenir toutes les unités lexicales simples existant dans d'autres modules, tels que le lexique-grammaire et le dictionnaire des mots composés. Par ailleurs, c'est le dictionnaire consulté lors de tout traitement impliquant la reconnaissance des mots d'un texte. DELAF est donc très central, mais comme les formes fléchies sont engendrées à partir des mots simples contenus dans le DELAS, ce dernier dictionnaire est au coeur du système. Il servira donc d'exemple pour présenter quelques caractéristiques des dictionnaires électroniques.

3.1. NOMENCLATURE D'UNITÉS FORMELLES

Dans l'élaboration de tout dictionnaire, le premier point à fixer est la définition des entrées. Pour construire le DELAS, dictionnaire de mots simples, il est donc nécessaire de préciser ce que nous entendons par mot ou unité simple, et d'en donner une définition adaptée à l'ordinateur, c'est-à-dire formelle.

Sous l'angle formel, une unité linguistique simple est un mot de la langue dont la forme écrite se présente comme une suite de caractères alphabétiques sans séparateurs. Ceci implique a priori la définition des caractères alphabétiques et des séparateurs : en première approche, l'alphabet constitutif des

mots simples a été limité à un seul type de caractères, les lettres minuscules, non accentuées et accentuées. Les séparateurs (trait d'union, apostrophe, espace blanc) sont exclus de cet alphabet.

Donc, par définition, un mot constitué de ce seul type de caractères, sans séparateur, est un **mot formel simple**, et doit être mis dans le dictionnaire DELAS. Par contre, un mot construit avec des mots simples et des séparateurs est un **mot formel composé**, et appartient au dictionnaire DELAC. Enfin tout mot incluant d'autres caractères : lettres majuscules, chiffres ou autres signes non alphabétiques est consigné dans des listes spéciales, à l'extérieur du DELAS et du DELAC.

La séparation des unités de la langue sur une base formelle est dépourvue d'ambiguïté, et elle correspond au découpage automatique des unités simples d'une phrase. Elle est essentielle pour l'analyse syntaxique. Mais la distribution des entrées dans le dictionnaire ne correspond pas aux critères usuels des lexicographes. En effet,

- elle intègre comme mot simple des juxtapositions plus ou moins complexes comme *autoradio* ou *désoxyribonucléique*,
- des entrées de mots traditionnellement inséparables telles *prud'homme*, *tohu-bohu*, *chemin de fer*, sont exclues du DELAS (Ces mots sont renvoyés au dictionnaire DELAC),
- par contre, les éléments de ces composés : *prud*, *tohu*, *bohu*, sont des mots formels simples et font partie des entrées du DELAS,
- les mots grammaticaux élidés (*c'*, *d'*, *j'*,...) sont aussi des mots d'entrée,
- de même les préfixes séparables par un trait d'union (*anti-*, *néo-*, *euro-*,...), et les constituants de locutions figées (en *catimini, ad hoc*,...).

Pour l'ordinateur, ces unités sont équivalentes à n'importe quel autre mot d'entrée. Elles servent de clé d'accès aux informations codées qui les accompagnent.

3.2. COUVERTURE LEXICALE ÉTENDUE

L'obligation pour les dictionnaires électroniques de disposer d'une couverture lexicale aussi étendue que possible a

déjà été soulignée. Ceci ne signifie pas pour autant que ces dictionnaires doivent accepter tout mot nouveau sans justification linguistique sérieuse.

A priori, nous avons choisi comme références pour le DELAS l'un et/ou l'autre des dictionnaires suivants :

- GDEL : Grand Dictionnaire Encyclopédique Larousse, 1982
- GR : Grand Robert de la Langue Française, 1986
- LX : Dictionnaire de la langue LEXIS, Larousse, 1979.

Tout mot trouvé dans l'une de ces sources est considéré comme attesté. Le Trésor de la Langue Française (TLF) est consulté pour des attestations ponctuelles. Des mots dont l'usage est signalé vieilli (notamment dans GR) sont aussi acceptés, mais avec une marque qui permet de les isoler éventuellement.

Dans la version actuelle du DELAS, les entrées lexicales couvrent non seulement la totalité des mots simples de la langue courante, mais aussi un grand nombre de termes techniques et de spécialités. On verra qu'une hiérarchisation des entrées a été adoptée afin de séparer le vocabulaire en différentes couches lexicales.

L'analyse automatique de gros corpus à l'aide du logiciel INTEX a permis d'augmenter sensiblement la couverture lexicale. En effet, après toute analyse effectuée, les mots non reconnus dans le dictionnaire DELAF sont examinés un à un. Puis, s'ils sont justifiés soit par la morphologie, soit comme néologisme, leur forme canonique est ajoutée au vocabulaire du DELAS. En fait, cette opération n'est effectuée qu'après un jugement favorable d'acceptabilité par au moins deux linguistes de l'équipe du LADL.

Beaucoup de mots nouveaux sont, nous l'avons remarqué, des dérivations ou des formations préfixées, pour lesquelles les médias se montrent plutôt créatifs. Par exemple on a pu relever dernièrement de nombreuses occurrences de formations venues du verbe *privatiser* :

privatisable, *imprivatisable,*
déprivatiser, déprivatisation, déprivatisable,

reprivatiser, reprivatisation, reprivatisable.

Toutes ces formations ont été introduites dans la dernière édition du DELAS.

Un programme se chargeant de la reconnaissance des dérivés d'une même famille, formés régulièrement à partir d'un mot de base, mais non explicitement présents en entrée de dictionnaire, a fait l'objet d'un travail de thèse de David Clémenceau. A ce jour ce traitement n'est pas encore implanté dans le système INTEX.

3.3. CODIFICATIONS SYSTÉMATIQUES

Un dictionnaire électronique, comme un dictionnaire d'usage, est constitué de mots d'entrées et d'informations associées. Ces dernières ne sont pas des définitions, mais sont des indications codées et formatées en vue de l'analyse syntaxique. Le caractère systématique de la codification est une propriété du dictionnaire électronique.

Dans le DELAS, les informations codées sont les suivantes :

- un numéro de couche lexicale (voir paragraphe 3.4),
- un code morphologique, composé d'un sigle de catégorie grammaticale et, pour les mots variables, d'un numéro de classe renvoyant à une liste des flexions de la langue :

  dormir,"1**.V26U**(i)
  succès,"1**.N2**
  monstrueux,"1**.A63**
  narcissiquement,"2**.ADV**
  parmi,"1**.PREP**

- une indication de la composition (quand l'entrée est une partie non-autonome de mot composé) :

  méli,"2.XN**{~-mélo}**(NN)
  pong,"2.XN**{ping-~}**(NN)

- des liens au lexique-grammaire des verbes (liens créés par le nom d'identification des tables syntaxiques où figure chaque verbe) :

  investir,"1.V18(t;**32L;37M;38LD;38L1**)

participer,"1.V3U(i;**33;35R;8**)

- des marques sémantiques (Hum, Anl, Conc,...) pour les substantifs :

linguiste,"1.N31**(Hum)**

perdrix,"1.N22**(Anl)**

- enfin des sous-marques informatives [Min, Vég,...] :

molybdénite,"3.N21(Conc)**[Min]**

jacinthe,"1.N21(Conc)**[Vég]**

sous-marques mises comme les marques d'usage (populaire, vieilli, argotique) en tant que commentaires, et non en tant que données utilisées par l'analyseur lexical.

Les difficultés inhérentes à la classification des mots, donc à leur codification, sont nombreuses. Les obstacles rencontrés vont de pair avec les problèmes de cohérence évoqués plus loin.

3.4. DISTRIBUTION DU LEXIQUE EN COUCHES

La répartition des entrées lexicales du DELAS en trois couches résulte d'un travail de thèse de doctorat de Mylène Bonan-Garrigues. Le numéro de couche attribué à chaque mot repose sur deux critères : la représentation et la plausibilité.

La représentation est une "image mentale - précise ou vague, forte ou faible, émotionnellement positive ou négative - signalant la simple présence du mot dans le lexique mental d'un individu". Partant de cette définition, M. Garrigues a pu séparer le lexique en mots connus, placés en couche lexicale 1, mots partiellement connus ou reconstitués, placés en couche lexicale 2, et mots inconnus, mis en couche 3.

Le critère de plausibilité a servi de test complémentaire. Il a consisté à évaluer intuitivement les possibilités d'apparition du mot dans l'ensemble des énoncés écrits ou oraux traités habituellement. Ainsi les mots *salsepareille*, *géhenne*, jugés moins plausibles que *chemin*, *définir*, *arbitraire*, sont placés dans une couche de numéro supérieur.

La première couche rassemble donc des mots courants du vocabulaire général, et de plausibilité forte. Il faut cependant noter que l'indice de plausibilité dépend du groupe de population visée et de la nature des textes soumis à l'analyse automatique. Ainsi les lexiques de la chimie, de la physique, de la médecine, du droit, des techniques, des arts, hautement spécialisés pour un profane et appartenant à la couche 3 du DELAS, pourraient être ramenés en couche 1 pour le spécialiste du domaine.

Dans l'état actuel de hiérarchisation du DELAS, près de 30% des entrées sont en couche 1, soit environ 24.000 mots, 18% sont en couche 2 et 52% en couche 3. Cette sélection comporte des limites, car la séparation des sens dans les entrées homographes est encore à l'étude. Des réajustements s'imposent au fur et à mesure des analyses de textes effectuées par le système INTEX. D'ores et déjà, nous avons pu vérifier que la confrontation des couches lexicales avec des textes techniques permet d'extraire le vocabulaire technique spécifique. Cette application est spécialement intéressante pour la documentation automatique professionnelle et pour l'établissement de lexiques thématiques.

3.5. COHÉRENCE DES STRUCTURES

L'élaboration d'un système tel que le DELA, composé de modules séparés, suppose de nombreux contrôles de cohérence. D'une part l'homogénéité interne de chaque module doit être assurée. D'autre part les données présentes dans les différents modules doivent être rigoureusement cohérentes entre elles.

A l'intérieur d'une structure, qu'il s'agisse de mots simples, de mots composés ou de tables syntaxiques, la codification systématique entraîne constamment des prises de décision basées sur des règles fixées au préalable. Les critères d'introduction des entrées, d'une part, et de codification, d'autre part, doivent rester stables et invariables tout au long de la construction du dictionnaire.

En ce qui concerne le DELAS, la cohérence interne repose sur plusieurs conditions, que nous allons illustrer par des exemples.

a - Entrées lexicales homogènes

Traditionnellement dans les dictionnaires, les noms de familles, animales ou végétales, sont entrés au pluriel (*cétacés, renonculacées*). Mais on trouve des attestations fréquentes d'emploi singulier, ce qui nous a mené à entrer ces substantifs au singulier dans le DELAS. Cette règle, une fois établie, doit être respectée sur toute l'étendue du dictionnaire, afin de maintenir l'homogénéité des entrées.

b - Familles de mots complètes et homogènes

A partir du terme *sympathie*, considérons la famille constituée du nom, de l'adjectif et de l'adverbe dérivés :

sympathie, *sympathique*, *sympathiquement.*

La famille analogue formée avec *empathie* sera homogène si sont présents les trois termes :

empathie, *empathique*, *empathiquement*.

On ne saurait cependant engendrer les familles de mots par une procédure automatique puisque le verbe *sympathiser* existe, mais non **empathiser*.

c - Codifications homogènes

L'adjectif *abrasif* a un emploi en nom masculin, réduction du composé NA : *un produit abrasif*. Il a donc deux codes dans le DELAS :

abrasif, .N1 .A38

Par comparaison, tous les adjectifs employés comme nom par une réduction de NA doivent avoir deux codes, tel :

calmant, .N1 .A32

d - Représentation de tous les emplois

En principe la mise des codes grammaticaux d'un mot ne pose pas de problème dans la mesure où il est monosémique et ne donne pas lieu à une grande variété d'emplois syntaxiques. Mais les mots polysémiques et d'emplois multiples ne sont pas si simples à coder. On connaît bien les phénomènes de passage d'une catégorie à une autre, tels que :

- les verbes et adjectifs substantivés
  le vivre et le couvert,
  l'utile et l'agréable,
- les réductions de composés nom/adjectif
  un tonique = un produit tonique,
  un généraliste = un médecin généraliste,
- les mots invariables employés comme substantifs
  le pourquoi, *le comment*.

Or, si tous ces emplois ne sont pas examinés en détail, et traduits par des codes adéquats pour chacun d'eux, le dictionnaire électronique reste avec des lacunes qui risquent d'être fatales à la reconnaissance. Donc, seul le codage complet de chaque entrée peut assurer la cohérence interne du module et sa conformité aux "exigences d'explicite" requises par l'ordinateur.

La cohérence globale concerne l'ensemble des dictionnaires dans leur relation les uns avec les autres. La cohésion du système est vérifiée par la confrontation des modules deux à deux :

DELAF / mots composés du DELAC
DELAS / lexique-grammaire
DELAS / exemples de phrases
DELAF / expressions figées.

La comparaison de ces modules par programmes permet de vérifier la concordance sur le plan des entrées lexicales et sur celui des informations codées.

Des ajustements sont nécessaires dès qu'une discordance est relevée. En pratique, toute modification des entrées d'une table syntaxique se répercute dans le DELAS et, par voie de

conséquence, dans le dictionnaire des formes fléchies du DELAF. Nous voyons apparaître ici les problèmes de **maintenance** d'un système de modules présentant des liens entre eux et en constante évolution.

CONCLUSION

L'élaboration de dictionnaires électroniques est à but informatique. Les applications se situent autant dans le domaine des recherches linguistiques que dans celui des "industries de la langue" : vérification orthographique, recherche documentaire, indexation de textes, études de concordances, analyse syntaxique.

Le caractère de quasi-exhaustivité des descriptions, exigé par l'ordinateur, offre l'avantage de mener à la collecte de données en grand nombre, et donc de permettre l'exploration de phénomènes lexicaux, non pas sur des exemples ou des listes établies manuellement, mais à partir de la totalité des occurrences présentes dans le dictionnaire ou dans de gros corpus.

Le système de description de la langue française élaboré au LADL s'applique à d'autres langues. Des dictionnaires électroniques sont déjà construits, ou en cours de réalisation, pour l'italien, l'anglais, l'allemand, l'espagnol, le portugais, le grec, le roumain, le polonais. Le principe de base étant celui de descriptions formelles et systématiques suivant le schéma du système DELA, les programmes d'applications sont indépendants de la langue traitée. En particulier, le logiciel d'interrogation et analyse de textes par INTEX a déjà été testé sur plusieurs langues de la communauté européenne.

BIBLIOGRAPHIE

BONAN-GARRIGUES Mylène. 1993 Méthodes de paramétrage des dictionnaires et grammaires électroniques. Thèse de doctorat, Université Paris 7

BOONS Jean-Paul, GUILLET Alain, LECLERE Christian. 1976 La structure des phrases simples en français. I Constructions intransitives. Genève-Paris, Droz

BOONS Jean-Paul, GUILLET Alain, LECLERE Christian. 1976 La structure des phrases simples en français. II Constructions transitives. Rapport de recherches du LADL, n˚6

COURTOIS Blandine, SILBERZTEIN Max. 1990 Dictionnaires électroniques du français. Langue française n˚87, Paris, Larousse

COURTOIS Blandine. 1990 Dictionnaire grammatical inverse du français : DELAS-I V06-2. Rapport technique du LADL n˚23

COURTOIS Blandine. 1990 Dictionnaire par classes flexionnelles : DELAS-CV, DELAS-CN, DELAS-CA version V06-2. Rapport technique du LADL n˚26

COURTOIS Blandine. 1992 Dictionnaire électronique des mots simples du français : DELAS V07-E1. Rapport technique du LADL n˚33

DANLOS Laurence. 1988 Les expressions figées. Langages n˚90, Paris, Larousse

DUBOIS Jean et al. 1973 Dictionnaire de linguistique. Paris, Larousse

GREVISSE Maurice, GOOSSE André. 1986 Le bon usage. Paris-Gembloux, Duculot

GROSS Gaston, JUNG René, MATHIEU-COLAS Michel, VIVES Robert. 1990 Le dictionnaire électronique des mots composés. Rapport technique du LADL.

GROSS Gaston. 1988 Degré de figement des noms composés. Langages n˚90, Paris, Larousse

GROSS Maurice. 1975 Méthodes en syntaxe. Paris, Hermann

GROSS Maurice. 1968 Grammaire transformationnelle du français. 1- Syntaxe du verbe. Paris, Cantilène

GROSS Maurice. 1986 Grammaire transformationnelle du français. 2- Syntaxe du nom. Paris, Cantilène

GROSS Maurice. 1990 Grammaire transformationnelle du français. 3- Syntaxe de l'adverbe. Paris, ASSTRIL

GIRY-SCHNEIDER Jacqueline. 1978 Les nominalisations en français. Genève, Droz

GUILLET Alain, LECLERE Christian. 1981 Formes syntaxiques et prédicats sémantiques. Langages n˚63, Paris : Larousse

GUILLET Alain, LECLERE Christian. 1992 La structure des phrases simples en français : constructions transitives locatives. Genève-Paris : Droz

KLARSFELD Gaby. 1990 Dictionnaire des anagrammes du français. Rapport de recherches du LADL.

LAPORTE Eric. 1988 Méthodes algorithmiques et lexicales de phonétisation de textes. Applications au français. Thèse de doctorat, Université Paris 7

LAPORTE Eric, SILBERZTEIN Max. 1989 Vérification et correction orthographiques assistées par ordinateur. Actes de la 1ère conférence européenne sur les techniques et applications de l'intelligence artificielle en milieu industriel. Hermès

LECLERE Christian. 1990 Organisation du lexique-grammaire des verbes français. Langue française n°87, Paris, Larousse

MAUREL Denis. 1989 Reconnaissance de séquences de mots par automates. Adverbes de date du français. Thèse de doctorat. Université Paris 7

MEUNIER Annie. 1981 Nominalisations d'adjectifs par verbes supports. Thèse de doctorat, Université Paris 7

PIOT Mireille. 1988 Conjonctions de subordination et figement. Langages N°90, Paris, Larousse

SILBERZTEIN Max. 1989 Dictionnaires électroniques et reconnaissance lexicale automatique. Thèse de doctorat, Université Paris 7

SILBERZTEIN Max. 1990 Dictionnaires électroniques et analyse automatique de textes. Le système INTEX. Paris, Masson, collection Informatique Linguistique.