Meta
Journal des traducteurs
Translators' Journal

META

Un dictionnaire électronique pour la reconnaissance des formes dérivées

Viviane Clavier et Muriel Coret

Volume 42, numéro 2, juin 1997

Lexicologie et terminologie II (1) et Traduction et post-colonialisme en Inde
Translation and Postcolonialism: India (2)

URI : https://id.erudit.org/iderudit/002337ar
DOI : https://doi.org/10.7202/002337ar

Aller au sommaire du numéro

Éditeur(s)

Les Presses de l'Université de Montréal

ISSN

0026-0452 (imprimé)
1492-1421 (numérique)

Découvrir la revue

Citer cet article

Clavier, V. & Coret, M. (1997). Un dictionnaire électronique pour la reconnaissance des formes dérivées. *Meta*, *42*(2), 307–316. https://doi.org/10.7202/002337ar

Résumé de l'article

Cet article présente un modèle de dictionnaire morphologique qui a été appliqué à la description d'environ 8 000 mots. Les auteurs soulignent les différents points de la description qui demandent à être développés et les acquis incontestables : l'intérêt que représente une description du lexique en termes de niveaux, qui structure le lexique et lui confère une hiérarchie propre à décrire les familles dérivationnelles.

UN DICTIONNAIRE ÉLECTRONIQUE POUR LA RECONNAISSANCE DES FORMES DÉRIVÉES*

VIVIANE CLAVIER ET MURIEL CORET
Université de Grenoble 3 et Université de Paris 7, France

Résumé

Cet article présente un modèle de dictionnaire morphologique qui a été appliqué à la description d'environ 8 000 mots. Les auteurs soulignent les différents points de la description qui demandent à être développés et les acquis incontestables : l'intérêt que représente une description du lexique en termes de niveaux, qui structure le lexique et lui confère une hiérarchie propre à décrire les familles dérivationnelles.

Abstract

This article presents a model of a morphological dictionary applied to the description of some 8000 words. The authors focus on various problems requiring further research as well as a positive contribution of the research: a lexical description in terms of levels is valuable in that it organizes the lexicon and provides a hierarchy that allows for describing derivational series.

INTRODUCTION

L'objectif est de poser les fondements nécessaires à l'élaboration d'un dictionnaire électronique (structure, contenu, description des entrées lexicales, exploitation des données), permettant de décrire la structure morphologique des mots dérivés par suffixation du lexique français. Un tel dictionnaire, exploitable par des programmes informatiques, permettra en reconnaissance de décomposer convenablement les mots suffixés du français ou, en génération, de produire tout mot dérivé bien suffixé de la langue.

Dans cet exposé, nous nous intéressons à l'organisation d'un dictionnaire morphologique sans spécifier les applications possibles, qui peuvent être la traduction automatique, la correction orthographique, la recherche d'informations (pour le regroupement de familles dérivationnelles autour d'un lemme de référence)...

Nous nous situons en effet en amont des applications puisque, quelles qu'elles soient, des décisions sont à prendre concernant l'analyse du lexique et les conséquences qu'elle entraîne sur le contenu du dictionnaire. Cette conception se rapproche de la tendance actuelle consistant à envisager la construction du dictionnaire électronique indépendamment des utilisations spécifiques qui en seront faites par les différentes applications[1]. Le dictionnaire dont il s'agit est donc, à proprement parler, une base de données d'où sont dérivés différents produits, répondant aux besoins précis des divers utilisateurs plutôt qu'un objet directement accessible à l'utilisateur final.

Le dictionnaire morphologique que nous envisageons est le produit d'une réflexion théorique sur ce qu'on entend par «mot suffixé» et sur la manière dont on peut le décrire. On se situe donc à l'intersection de problèmes de description linguistique et de modélisation TAL.

La réflexion que nous présentons ici est le résultat d'une mise en commun de travaux menés par M. Coret et V. Clavier, respectivement à Paris 7 et Grenoble 3.

Le travail de Coret (1994) s'appuie sur un corpus de mots en *-age, -ment, -ion, -eur*, et a pour objet la segmentation des mots en base et suffixe, précédé éventuellement d'un «segment d'allongement» (le -at- de *programm-at-ion*). Cette analyse isole une unité morphologique particulière, le segment d'allongement, qui joue un rôle classificatoire en synchronie dans l'organisation du lexique. Elle distingue des classes de bases et de suffixes et fixe les premières conditions de leur association.

Le travail de Clavier (1995) traite également de segmentation et d'interprétation des mots dérivés. Il s'appuie sur un corpus de mots en *-age, -aire, -eur, -ier, -ment, -tion*, dans une perspective TAL, l'objectif étant de construire un dictionnaire morphologique accessible à un programme d'analyse dans le système de reconnaissance du français écrit CRISTAL. La réflexion sur les unités minimales et la nature des informations à prendre en compte dans le dictionnaire fait émerger une hiérarchie des unités (morphèmes, lemmes, métalemmes) à laquelle fait écho l'organisation même du dictionnaire.

Par-delà les motivations de chacune et les divergences qui ont émergé lors de certaines analyses, il s'est dégagé des points de consensus, que nous voulons mettre ici en avant. Ces points d'accord portent en particulier sur la conception du lexique et le contenu qui en résulte pour le dictionnaire, ainsi que sur la définition du mot suffixé.

Nous nous intéressons ici dans un premier temps à des problèmes théoriques de description des mots dérivés, questions qui se posent (ou qui devraient se poser) de toute façon lors de l'élaboration d'un dictionnaire, quel que soit le cadre envisagé. Dans un deuxième temps, nous présentons un modèle d'organisation de dictionnaire inspiré du dictionnaire morphologique pour le système CRISTAL.

1. MORPHOLOGIE DÉRIVATIONNELLE : PARTIS PRIS THÉORIQUES

1.1. Le lexique

L'existence et la nécessité du lexique comme composante de la grammaire sont maintenant bien établies. Mais la définition qui en est donnée est sujette à controverse. Considéré comme produit d'entrée de la composante syntaxique, le lexique est souvent défini comme une liste de formes qui sont, sur le plan morphologique, indifféremment simples ou construites. Figureront au même titre des formes simples (*laver*), des formes dérivées (*prélavage*), des formes composées (*lave-linge*), des mots-valises (*lavomatic*)... Il en résulte un lexique hétérogène qui présente le double inconvénient d'être redondant — il ne distingue pas le simple du construit — et non hiérarchisé — il met sur le même plan le régulier et l'irrégulier.

Cette position comporte enfin l'inconvénient de réduire le lexique à une liste, nécessairement close à un moment donné, ce qui ne rend pas compte de la créativité lexicale.

Nous voudrions au contraire définir un lexique non redondant, constitué d'unités élémentaires. Ce lexique sera l'entrée de règles morphologiques dont le rôle est de produire le lexique possible de la langue, non fini par définition. Ce sont ces règles qui prennent en charge la hiérarchisation des unités, en traitant l'opposition entre le régulier (produit par la règle, non consigné) et l'irrégulier (non prédictible, consigné).

Nous reprenons en fait la distinction entre lexique restreint et lexique étendu telle qu'elle est établie dans Fradin (1993b: 14-15) à la suite de Zwicky (1992). Elle oppose un lexique défini en intension, «le noyau du lexique dénué de toute redondance pouvant servir de base à l'inventaire complet de lexèmes de la langue si on se donne un ensemble approprié de règles morphologiques» à un lexique défini en extension, qui comporte «l'inventaire total, mais possiblement infini, des lexèmes».

Le dictionnaire qui consigne cette réalité du lexique comporte les unités simples qui apparaissent de manière récurrente dans d'autres unités de la langue. Ces unités simples appartiennent au lexique restreint, défini en intension.

Cette stratégie de calcul des formes, qui consiste à ne retenir que les unités élémentaires accompagnées d'une batterie de règles, est la seule adéquate pour répondre au problème de l'optimisation de la couverture lexicale. Sur le plan de la langue comme sur celui des applications TAL, cette manière de décrire le lexique est la seule possible si l'on veut traiter la néologie. Sa mise en œuvre n'est cependant pas triviale, et rares sont les systèmes qui ont recours à cette stratégie.

Le lexique étendu est donc le produit de règles lexicales fort différentes (dérivation, composition...) : nous restreignons ici notre description au lexique des formes dérivées par suffixation.

1.2. Le mot suffixé

Nous adoptons une perspective morpho-sémantique de la dérivation, selon laquelle les unités construites sont décomposables en constituants minimaux formellement et sémantiquement identifiables. Cette perspective s'inscrit dans le courant établi par Dell (1970), Halle (1973), Aronoff (1976) et Corbin (1987) et postule que le sens d'un mot construit est compositionnel par rapport à sa structure. Cela suppose que le sens à considérer est le sens prédictible, induit par la structure, et non le sens attesté tel qu'il se donne à voir dans les dictionnaires de langue usuels. Au-delà de la diversité des emplois d'unités telles que *câblage, cerclage, aconage*[2], par exemple, nous cherchons à mettre en évidence la structure commune à ces trois unités, dont on rend compte par le fait qu'elles comportent le même suffixe et que, à un certain niveau de généralité, leur interprétation est similaire. Le dictionnaire usuel, qui ne reflète pas cette réalité, ne propose pas systématiquement une paraphrase dérivationnelle et s'attache souvent davantage à la description du référent, objet du monde, qu'à celle de l'objet linguistique. Derrière la diversité et l'hétérogénéité, nous cherchons le sens prédictible, qui n'est pas un donné immédiat.

Soit l'exemple de *croupier*, dont les définitions dans le Littré sont dans Clavier (1995 : 1re partie, Ch. III) :

> **croupier (1)** : s.m. **1.** Celui qui est associé avec le joueur tenant la carte ou le dé. [...]. **2.** Anciennement, celui qui assistait le banquier à la bassette et l'avertissait des cartes qu'il passait. **3.** Celui qui prend part à quelque affaire de finance, sans s'y faire nommer. **4.** Terme de droit canonique. Confidentiaire prêtant son nom à celui qui plaide pour un bénéfice. **5.** Adjectivement. Qui est en croupe, derrière une personne en selle.
>
> **croupière (2)** : s.f. **1.** Partie du harnais qui, passant par-dessous la queue du cheval, vient se rattacher à la selle par-dessus la croupe. **2.** Terme de marine. Grelin qui, attaché par un bout au câble près de l'ancre avant de la mouiller, passe de l'autre bout par l'un des sabords de l'arrière. **3.** s.f. plur. Pièces qui tiennent en état le devant ou le derrière d'un train de bois.

Le fait que le mot soit utilisé pour désigner des référents aussi différents que «employé de maison de jeu», «homme d'affaires», «partie du harnais du cheval», «grelin» et «pièces d'un train de bois» n'est pas un obstacle à son analyse morphologique puisqu'on peut déterminer un sens structural commun que l'on peut paraphraser en «qui se trouve à la croupe, *i.e.* à l'arrière, en retrait de quelque chose».

Parallèlement à ces distorsions, la forme peut également subir un certain nombre d'altérations, calculables ou non, qui peuvent brouiller la reconnaissance immédiate des constituants. Ainsi, on reconnaîtra dans *absorption* la même base que dans *absorber* et on rapprochera également *strangulation* de *étrangler*, *promotion* de *promouvoir*, *extinction* de *éteindre*... même si ces derniers cas ne relèvent pas d'une règle phonologique générale du français.

Différents mécanismes aussi bien au niveau du sens que de la forme sont à mettre en place pour traiter ces différents types de distorsions, comme l'a montré Corbin (1987)[3].

1.3. La segmentation en unités minimales

Dans une analyse morphologique, on est forcément confronté à des questions de segmentation et d'interprétation des unités repérées. En effet, dans un premier temps, les segmentations s'opèrent sur des critères qui sont essentiellement distributionnels et tout le problème est de donner un statut linguistique aux segments obtenus.

La première étape consiste donc à segmenter une forme en base et suffixe. Les dictionnaires et les grammaires usuels font apparaître à ce propos des divergences très nettes quant au nombre et à la forme des suffixes du français (Coret 1994). Pour une large part, ce qui préside aux procédures de segmentation est la volonté de simplifier et d'homogénéiser la description des suffixes du français. Le parti pris est de lier la décomposition à l'apparition récurrente de segments terminaux : nous identifions ainsi un suffixe déverbal *-age*, qui n'a qu'une seule réalisation, rejetant les analyses qui font de *-age* et *-issage* des allomorphes. Nous posons donc une forme unique *-age*, qui se construit sur des bases liées à des verbes. Lorsque ceux-ci sont du deuxième groupe, la suffixation prend pour base le radical de l'imparfait — ce qui rend compte de la présence de *-iss-*. Il en va de même d'ailleurs pour *-ment* et *-eur* (féminin *-euse*). Nos analyses aboutissent à des segmentations communes dans la majorité des cas.

Cependant, pour les noms et adjectifs en *-(at / it / t)-ion, -(at / it / t)-eur* et *-(at / it / t)-if*, qui sont des cas plus délicats, nos conclusions divergent. Ainsi dans la liste :

programmeur,
programmateur,
programmer,
créateur,
créatif,
créer, etc.

une segmentation formelle isole un segment *-at-*. On a *a priori* trois unités : *programm-at-eur, cré-at-if*... mais c'est une analyse plus fine qui doit permettre d'affecter à ce segment un statut linguistique. On a en effet le choix entre deux solutions : le rattacher à une unité déjà connue, la base ou le suffixe ou créer une nouvelle unité, le «segment d'allongement» (cf. pour une première analyse du «joncteur», Gruaz 1988). La première analyse s'exprime en termes d'allomorphie : c'est la plus courante. Ces deux positions font encore l'objet de débats au sein de la communauté linguistique (cf. Plénat 1988 ; Gruaz 1988 ; Fradin 1993a et Huot 1994). Dans tous les cas, quelle que soit la solution retenue, la question du conditionnement de l'apparition de ce segment se pose. Il est un fait que *-at-* ne peut pas apparaître devant n'importe quel suffixe, ni derrière n'importe quelle base. Mais la détermination du conditionnement qui sera proposé ne permettra pas de trancher sur le caractère autonome ou non de ce segment. En effet, cette démarche revient toujours à chercher ce qui, dans la base, déclenche l'apparition du segment en termes de propriétés morphophonologiques — repérage de suites consonantiques particulières (str *(démonstrateur, strangulation)*, sp *(aspirateur, spoliation)*, sph *(blasphémation)*, etc.), et implication des bases dans des paradigmes flexionnels du premier, deuxième ou troisième groupe. L'enjeu d'un tel travail n'est pas moindre ; cela permettrait de redéfinir en synchronie l'opposition populaire / savant, qui est énoncée en termes d'histoire de la langue.

En revanche, le choix entre allomorphie et segment d'allongement est lié à l'analyse qu'on en fait sur le plan interprétatif. Pour Huot (1994 : 55) en effet, «[...] on peut (également) reconnaître à ce suffixe thématique VC un certain contenu interprétatif. Mais à la différence des suffixes dérivationnels, dont l'interprétation, sans doute large, est néanmoins caractérisable («action de», «agent de», «qui a la capacité de», etc.), la valeur de ce suffixe serait, et c'est là l'important, seulement aspectuelle et n'exprimerait rien d'autre que l'accompli. Ainsi un *format* n'est que ce qui résulte de l'effectuation de l'acte de mettre

en forme, une *faillite* désigne la situation résultant de la défaillance effective (d'une entreprise ou d'un commerçant), etc.»[4]. Attribuer à ce segment une valeur sémantique incite à lui accorder un statut à part entière et éliminerait l'hypothèse de l'allomorphie.

Malgré des conclusions différentes (pour Coret (1994), il s'agit d'un segment autonome, pour Clavier (1995), d'une allomorphie suffixale), il y a accord sur la forme des bases. Que le segment *-at-* soit intégré au suffixe ou non, son apparition sera prise en charge par des règles dérivationnelles.

1.4. Classification des unités minimales

Après avoir procédé à la segmentation des unités construites, nous nous intéressons aux propriétés des unités obtenues, suffixes et bases, et à la façon d'en rendre compte dans le couple dictionnaire-grammaire.

Nous nous situons dans le cadre généralement admis selon lequel les suffixes sont considérés comme des opérateurs qui agissent sur des opérandes, les unités lexicales, pour construire d'autres unités lexicales (Corbin 1992[5] ; Fradin 1993a). Dans cette optique, les opérateurs suffixaux sélectionnent un certain nombre de propriétés des unités lexicales qui dépendent de différents niveaux d'interprétation.

Une unité lexicale a en effet des propriétés phonologiques, graphiques, morphologiques, syntaxiques et sémantiques.

Ainsi, c'est parce que *strangul-* est réservé aux noms en *-ion* avec un segment d'allongement *-at-*, que **strangulement* est impossible. Cette séquence ne respecte pas les contraintes morphologiques que sélectionne le suffixe *-ment* : *strangul-* ne peut apparaître sans *-at-* et, d'autre part, l'apparition de ce segment est incompatible avec le suffixe *-ment*[6].

Parallèlement, c'est pour des raisons sémantiques que *brassière* ne peut pas être relié à *brasse* et doit être interprété en relation à *bras* : le suffixe *-ier* sélectionne en effet des bases nominales à valeur [+ concret], ce qui élimine la base nominale *brasse* qui renvoie à une activité — cf. Clavier et Lallich-Boidin (1994) qui renvoie à Corbin et Corbin (1991).

Nous n'illustrons pas ici toutes ces propriétés. Mais ces exemples suffisent à montrer que les règles de suffixation font intervenir différents types de propriétés des bases. Soulignons que l'autonomie syntaxique de la base ne figure pas parmi celles-ci. En effet, la forme *cécité* sera analysée de la même manière que *acidité* et *avidité*, même si *céc-* n'a pas d'autonomie dans la langue. On reconnaît dans ces trois formes le même suffixe qui construit des noms de qualité sur des adjectifs (pour l'interprétation de *céc-*, on a recours à la forme supplétive *aveugle*).

Ces différents niveaux d'interprétation des bases et les contraintes que font peser sur eux les suffixes de la langue doivent être consignés dans le dictionnaire et la grammaire. Ces informations donnent l'organisation du dictionnaire qui doit comporter différents niveaux de description et faire ressortir les relations entre les bases.

Ainsi, à la conception d'un dictionnaire présenté comme une liste de formes non hiérarchisées s'oppose un dictionnaire structuré dont les entrées sont le résultat d'une analyse morphologique. Il ne s'agit pas en effet de consigner telles quelles les séquences issues de la segmentation. Ainsi, les segmentations de *absorp-tion, absorb-er, truc-age* et *truqu-age* ne fournissent pas quatre entrées lexicales. On procède à des regroupements qui permettent de réduire le nombre d'entrées. Ainsi, on obtient trois niveaux de description :

- Au niveau morphologique : les séquences *truc-* et *truqu-* (*truc, trucage, truquer*) et *absorb-, absorp-* (*absorber, absorption*) sont ramenées respectivement à une unité lexicale unique *truqu-* et *absorp-*. En effet, il existe des règles phonographiques très générales en français qui conditionnent ces variations *qu / c* et *b / p* en fonction de la nature phonographique du contexte. Les règles qui décrivent les réalisations en surface de ces unités relèvent de la grammaire de la langue et ces variantes ne sont pas représentées dans le dictionnaire.

■ Au niveau syntaxique : les formes alternantes *produi(s)-*, *produc-* (*produire, production*) constituent des entrées indépendantes au niveau morphologique. En effet, contrairement au cas précédent, il n'existe pas de règle générale à appliquer pour rendre compte du passage de la forme *produi(s)-* à la forme *produc-*. Ces alternances concernent des listes finies d'unités (*condui(s)-, conduc-, indui(s)-, induc-...*). Pour les mêmes raisons, *mett-* et *miss-* (*mettre, mission*), *concéd-, concess-* (*concéder, concession*) seront regroupés.
Au niveau syntaxique, on aura une forme lemmatisée commune, désignée par la forme à l'infinitif pour les verbes, le singulier pour les noms et le masculin singulier pour les adjectifs : *produire* (V), *conduire* (V), *etc.*

■ Au niveau sémantique : certaines unités, distinguées aux niveaux morphologique et syntaxique, seront regroupées au niveau sémantique. Les unités *céc-* et *aveugle*, par exemple, constituent deux lemmes au niveau syntaxique, car la relation qui existe entre elles est d'ordre strictement sémantique et il n'existe aucun lien formel entre ces deux unités. Mais elles reçoivent une description sémantique unique qui rend compte de leur relation de parasynonymie. On parle classiquement de «supplétion lexicale» quand deux unités lexicales n'ont aucune relation formelle mais appartiennent au même paradigme flexionnel ou dérivationnel.

Ces différents niveaux de description rendent compte du statut à attribuer aux unités lexicales élémentaires constitutives du système. Ce ne sont pas des «mots», puisque l'autonomie syntaxique n'est pas retenue comme propriété nécessaire. Ce sont des unités définies sous la forme d'un triplet de propriétés sur lesquelles agissent les règles de suffixation.

1.5. Grammaire de la suffixation

Le travail de reconnaissance des unités minimales impliquées dans la dérivation est corrélé à l'élaboration d'une grammaire de la suffixation. Celle-ci s'exprime par un ensemble de règles dérivationnelles, qui prennent en compte les différents niveaux d'analyse cités ci-dessus. En ce qui nous concerne, nous nous sommes surtout intéressées à l'aspect des règles qui traitent de la forme (phonographique, morphologique), en reprenant pour le cadre général de leur formulation les travaux de Corbin (1987) et Fradin (1993a).

On peut représenter un opérateur suffixal comme un ensemble de contraintes formelles, catégorielles et sémantiques qui s'exercent sur les unités lexicales élémentaires et dérivées. Ces contraintes sont de deux types :

■ Des contraintes générales qui concernent l'ensemble des suffixes et qui déterminent par exemple la forme du radical verbal utilisé en dérivation (base de l'imparfait pour les verbes des deuxième et troisième groupes : *verniss-* dans *vernissage*, *buv-* dans *buveur*). Ces contraintes doivent être décrites dans une grammaire.
Ces règles donnent aussi certaines des conditions de réalisation des unités dérivées. Elles stipulent, par exemple, les cas où le suffixe *-ier* est réalisé *-er* et ceux où il apparaîtra sous la forme *-ier* : *-er* derrière *ch-* [ʃ], *g-* [g], *ill-* et *y-* [j] (dans *archer, boulanger, oreiller, papayer*) *versus -ier* partout ailleurs (*tablier, abricotier...*)[7].

■ Des contraintes particulières à chaque suffixe qui seront consignées dans le dictionnaire. Ces règles posent les conditions que doivent remplir les unités lexicales élémentaires et les suffixes en vue de leur association et fixent certaines des propriétés du mot construit (catégorie, sens prédictible). Ainsi, par les propriétés qu'elles mettent en jeu aussi bien pour l'unité lexicale élémentaire que pour l'unité lexicale dérivée, ces règles permettent de distinguer deux suffixes homonymes *-eur* : $-eur_1$ construit des noms d'agent à partir de verbes (*joueur*) et $-eur_2$, des noms de propriété à partir

d'adjectifs (*blancheur*). C'est aussi dans le dictionnaire que sera précisé si le suffixe requiert l'apparition d'un segment d'allongement.

Seul ce deuxième ensemble de contraintes est inscrit dans le dictionnaire des suffixes.

1.6. Cahier des charges

Pour couvrir l'inventaire possiblement infini des mots de la langue, on adopte une stratégie calculatoire afin de doter le système informatique chargé d'exploiter le dictionnaire d'une capacité prédictive. En écho à cette conception, on peut définir un dictionnaire dont les entrées seront les unités lexicales élémentaires auxquelles s'appliquent un ensemble de règles de construction d'unités lexicales. Nous nous sommes limitées à la description des mots suffixés, pour lesquels le sens et la forme sont le produit d'une certaine structure, le sens étant compositionnel par rapport à la forme, malgré les distorsions apparentes. Ces distorsions peuvent être prises en charge soit, de façon statique, par marquage dans le dictionnaire, soit, dynamiquement, par les règles dérivationnelles. Il est apparu que ces dernières font appel à des propriétés des unités lexicales (suffixes et bases) relevant de différents niveaux d'interprétation.

On peut établir le cahier des charges d'un dictionnaire morphologique du français comme suit :

- le dictionnaire est conçu pour être couplé à un analyseur morphologique (en reconnaissance) ou à un générateur (en production). Il est donc indépendant de l'application ;
- il est constitué des unités lexicales élémentaires de la langue, dont l'identification est le résultat d'une analyse morphologiquement justifiée ;
- il est structuré selon trois niveaux, morphologique, syntaxique, sémantique, et donne à voir une hiérarchie représentative des relations entre les unités lexicales (allomorphie, supplétion, parasynonymie).

On construit ainsi un dictionnaire homogène, non redondant et hiérarchisé.

2. MODÈLES DE DICTIONNAIRES MORPHOLOGIQUES POUR LE TAL

Conformément aux exigences évoquées précédemment, le dictionnaire doit présenter les différents niveaux de description des unités lexicales élémentaires. Par ailleurs, on peut construire un dictionnaire de suffixes qui s'appuie sur la grammaire dérivationnelle du français et qui donne pour chaque suffixe les conditions de son fonctionnement.

Les dictionnaires que nous présentons ici sous forme de prototypes reprennent, pour l'organisation de leur contenu (niveaux de description et types d'informations propres à chaque niveau), le dictionnaire CRISTAL établit, dans le cadre d'un programme de reconnaissance pour une application spécifique, la recherche d'informations. Les résultats de Clavier (1995) nous fournissent en effet un modèle d'organisation que nous illustrerons ici en nous concentrant sur les points d'analyse qui convergent avec ceux de Coret (1994).

2.1. Le dictionnaire des unités lexicales élémentaires

Le dictionnaire comporte les trois niveaux de description évoquée au paragraphe 1.4.

Le niveau morphologique regroupe les différents allomorphes (*absorb-* et *absorp-*) dont la distribution est prédictible et peut être décrite par des règles générales. Mais il ne réunit pas les formes alternantes (*produc-, produi(s)*).

Le niveau syntaxique comporte la forme lemmatisée et catégorisée des unités de niveau morphologique. On constate que notre définition du lemme n'est que partiellement

conforme à celle que l'on trouve dans les dictionnaires de langue : sur le plan de la forme, on trouve aussi bien des formes canoniques standard (infinitif pour les verbes, masculin ou féminin singulier pour les noms et adjectifs) que des formes non autonomes qui ne présentent pas de réalisation flexionnelle (*céc-*). À ce niveau, on regroupe les formes alternantes.

Le niveau sémantique regroupe entre elles les unités de niveau syntaxique en relation de parasynonymie et associe à chaque entrée une description sémantique dont la forme ultime reste à définir (paraphrase définitionnelle, complexe de traits, etc.).

On aura :

Forme de surface	**Niveau morphologique**	**Niveau syntaxique**		**Niveau sémantique**
absorber	absorb-	absorber	V	«laisser pénétrer et retenir un liquide dans sa substance»
absorption	absorb-	absorber	V	«laisser pénétrer et retenir un liquide dans sa substance»
aveugle	aveugle	aveugle	Adj	«qui est privé du sens de la vue»
cécité	céc-	céc-	Adj	«qui est privé du sens de la vue»
magasin	magasin	magasin	N	«lieu de dépôt de marchandises»
produire	produi(s)	produire	V	«produire, faire exister»
production	produc-	produire	V	«produire, faire exister»

2.2. Le dictionnaire des suffixes

Le format général d'une entrée suffixale peut être schématisé comme suit :

forme_suffixe

Contraintes formelles$_1$	→	contraintes formelles$_2$
Contraintes syntaxiques$_1$	→	contraintes syntaxiques$_2$
Contraintes sémantiques$_1$	→	contraintes sémantiques$_2$

Les contraintes formelles concernent les caractéristiques morphologiques des unités lexicales avec lesquelles les suffixes se combinent. Ce sont ces règles qui doivent indiquer que, entre les deux bases alternantes, *concess-* et *concéd-*, *-ion* sélectionne la première, et que, derrière la base *programm-*, *-ion* sera précédé du segment d'allongement *-at-*. Traditionnellement, ce conditionnement est traité par un marquage en lexique à l'aide du trait [± savant] attribué aux bases et aux suffixes. Cette solution présente un double inconvénient : elle utilise l'opposition mal définie savant/populaire ; le marquage en traits est une solution redondante qui n'a aucune capacité prédictive. Nous avons déjà mentionné

au paragraphe 1.3. la nécessité de définir cette opposition en synchronie. Cela permettrait d'abandonner la stratégie de marquage et de rendre compte du phénomène au niveau de la grammaire. Dans l'état actuel du dictionnaire, nous marquons cette information par la mention «sélection d'une base complexe» attribuée au suffixe.

Le niveau syntaxique note les informations sur la catégorie des unités lexicales. Il est évident qu'une telle information est insuffisante pour caractériser chaque suffixe, notamment pour le triplet des suffixes concurrents *-age, -ment, -ion.* Le recours à la sous-catégorisation [± transitif] paraît lui aussi insuffisant et doit vraisemblablement être complété par des informations sur la nature des procès mis en jeu[8].

Le niveau sémantique figure de manière symbolique dans ce modèle de dictionnaire. Sa mise en application nécessite une véritable théorie sémantique de la langue, et en particulier, pour les suffixes qui nous concernent, d'une sémantique des procès — qui reste à construire.

Voici à titre d'exemple les entrées lexicales que nous proposons pour les suffixes *-age*$_1$ (dénominal), *-age*$_2$ (déverbal), *-ment* (déverbal) et *-ion* (déverbal).

Suffixe	Contraintes formelles	Contraintes syntaxiques	Contraintes sémantiques
-age1	[- complexe] → [- complexe] *abatt-abattage*	V → N	«procès» → «action» implication d'un agent contrôleur
-age2	[- complexe] → [- complexe] *lait-laitage*	N → N	«objet» → «collectif»
-ment	[- complexe] → [- complexe] *abatt-abattement*	V → N	«procès» → «action» absence d'agent contrôleur
-ion	[- complexe] → [± complexe] *programm-programmation* *produc-production*	V → N	«procès» → «action»

CONCLUSION

Le modèle de dictionnaire que nous avons proposé a, concrètement, été appliqué à la description d'environ 8 000 mots dans Clavier (1995). Ce prototype doit prochainement faire l'objet d'une implémentation dans le cadre de CRISTAL, ce qui permettra de valider le modèle sur un corpus de textes.

Nous avons souligné au cours de cet exposé les différents points de la description qui demandent à être développés. Certains d'entre eux peuvent être traités à partir de nos corpus, en particulier la reformulation en synchronie de l'opposition savant/populaire. Mais les questions de segmentation ne se limitent pas à un travail sur la forme des unités. La détermination du statut linguistique du segment d'allongement, par exemple, passe inévitablement par le recours à une sémantique des procès, qui reste à construire.

Parmi les acquis incontestables, on peut souligner l'intérêt que représente une description du lexique en termes de niveaux, qui structure le lexique et lui confère une hiérarchie propre à décrire les familles dérivationnelles. Il est maintenant bien établi qu'en traitement automatique des langues, l'identification des liens sémantiques au sein du lexique est une étape incontournable, et la mise en évidence des structures morphologiques y contribue de manière décisive.

Notes

* Cet article est issu d'une communication présentée par l'auteur aux IV[es] Journées scientifiques du réseau «Lexicologie, terminologie, traduction» de l'AUPELF-UREF (Lyon, France, 28, 29, 30 septembre 1995).

1. C'est ce qui est ressorti nettement des propos tenus par concepteurs, constructeurs et utilisateurs de dictionnaires réunis à Bruxelles lors du Colloque «Analyse de la valeur des dictionnaires spécialisés», 31 mai et 1[er] juin 1995, CTB, Marie-Haps.
2. câblage : «1°. Fabrication d'un câble ; torsion des fils d'un câble (...)»
   cerclage : «Action de cercler».
   aconage : «*Mar.* Opération de chargement ou de déchargement d'un navire au moyen d'acons».
3. On se reportera également à *Lexique* 10, qui présente plusieurs travaux de l'équipe SILEX.
4. Remarquons que pour l'auteur, le *-at-* terminal de *format, agrégat* est le même que celui de *formation, agrégation...*
5. «[Un suffixe] est un opérateur sémantique servant à exprimer des relations ou des opérations mettant en jeu des unités linguistiques à pouvoir référentiel» (Corbin 1992 : 25-55).
6. On oppose en effet les suffixes *-ion, -eur* (féminin *-rice*) aux suffixes *-age, -eur* (féminin *-euse*), *-ment* : les premiers font apparaître un segment joncteur et pas les seconds, cf. Coret (1994).
7. Pour un exposé complet, voir Clavier (1995).
8. Dans la répartition *-age -ment*, la notion d'agent contrôleur définie dans Desclés (1993) paraît en effet centrale. Cf. l'opposition *abattage* (par un agent) versus *abattement* (sans agent). Voir Coret (1994) pour une illustration des limites d'une description en termes de [± transitif].

RÉFÉRENCES

ARONOFF, M. (1976) : *Word formation in Generative Grammar*, Linguistic Inquiry, Monograph one, Cambridge (Mass.), The MIT Press.

BERRENDONNER, A. (1995) : «Redoublement actantiel et nominalisations», *Scolia*, revue de l'Université de Strasbourg (à paraître).

CLAVIER, V. (1995) : *Modélisation de la suffixation pour le traitement automatique du français : Application à la recherche d'informations*, Thèse de doctorat, Grenoble.

CLAVIER, V. et G. LALLICH-BOIDIN (1994) : «Modélisation linguistique de la suffixation en vue de l'analyse automatique», *TAL*, 35 (2), pp. 129-143.

CORBIN, D. (1987) : *Morphologie dérivationnelle et structuration du lexique*, 2 volumes, Tübingen, Max Niemeyer.

CORBIN, D. (1992) : «Hypothèses sur les frontières de la composition nominale», *Cahiers de grammaire-Toulouse*, 17, pp. 25-55.

CORBIN, D. et P. CORBIN (1991) : «Un traitement unifié du suffixe -ier(e)», *Lexique*, 10, pp. 61-145.

CORET, M. (1994) : *Problèmes de suffixation et structuration du lexique. Étude des mots en -eur, -age, -ment, -ion*, Thèse de doctorat en linguistique, Université Paris 7.

DELL. F. (1970) : *Les règles phonologiques tardives de la morphologie dérivationnelle du français*, Thèse de doctorat, Cambridge (Mass.), The MIT Press.

DESCLÉS, J. P. (1993) : *Représentation des connaissances : archétypes cognitifs, chaînes conceptuelles et schémas grammaticaux*, Centre d'analyse et de mathématiques sociales, CNRS EHESS-Sorbonne.

FRADIN, B. (1993a) : *Organisation de l'information lexicale et interface morphologie / syntaxe dans le domaine verbal*, Thèse de doctorat d'État, Université Paris 8.

FRADIN, B. (1993b) : «La théorie morphologique face à ses choix», *Cahiers de lexicologie*, 63, pp. 5-42.

GRUAZ, C. (1988) : *La dérivation lexicale en français contemporain*, Rouen, PUR, n° 114.

HALLE, M. (1973) : «Prolegomena to a Theory of Word Formation», *Linguistic Inquiry*, IV (1), pp. 3-16.

HUOT, H. (1994) : «Sur la notion de racine», *TAL*, 35 (2), pp. 49-75.

LALLICH-BOIDIN, G., HENNERON, G. et R. PALERMITI (1990) : *Analyse du français. Achèvement et implantation de l'analyseur morpho-syntaxique*, Grenoble, Cahiers du CRISS, n° 16.

PLENAT, M. (1988) : «Morphologie des adjectifs en -able», *Cahiers de grammaire-Toulouse*, 13, pp. 101-132.

ROUAULT, J. (1987) : *Linguistique automatique. Applications documentaires*, Berne, Peter Lang.

ZWICKY, A. M. (1992) : «Some Choices in the Theory of Morphology», R. Levine, *Formal Grammar. Theory and Implementation*, Oxford, OUP.