La collecte et l'utilisation des données en morphologie

Nabil Hathout[*], Fiammetta Namer[**], Marc Plénat* & Ludovic Tanguy*

1. Introduction

Les progrès de l'informatique donnent accès à un très grand nombre de néologismes de toute sorte. Il est devenu aisé de constituer d'immenses listes de formes nouvelles rassemblant aussi bien des mots d'auteur ou des termes techniques que des mots de la langue courante. Le développement des capacités de stockage a en effet permis la constitution de très grandes bases de textes. Parallèlement, l'augmentation des capacités de calcul a rendu possible le traitement de ces très grandes quantités de données dans un temps acceptable. Les premiers corpus français de textes informatisés étaient essentiellement composés d'œuvres littéraires ; la base *Frantext* est sans doute le meilleur exemple de ce type de corpus en France. Les chercheurs ont ensuite disposé d'archives électroniques de journaux comme *le Monde* ou *le Soir*. Plus récemment, enfin, l'avènement de la Toile a fourni aux linguistes d'énormes quantités de textes très divers, allant des blogs et des forums jusqu'aux textes officiels et aux écrits littéraires ou scientifiques.

La communauté des morphologues dans son ensemble est assez peu préparée à tenir compte de ces possibilités nouvelles. L'utilisation des moyens informatiques n'entre que très lentement dans les mœurs ; on tient d'ordinaire pour acquis que le dictionnaire et l'intuition suffisent à séparer le bon grain des généralisations pertinentes de l'ivraie des formes agrammaticales. Il a fallu des circonstances particulières – la présence d'informaticiens-linguistes dans quelques laboratoires de linguistique comme l'ERSS ou l'ATILF – pour que les possibilités nouvelles soient exploitées et pour qu'il soit ainsi renoué en morphologie avec la tradition philologique qui fondait la linguistique sur la lecture critique des textes. A notre connaissance, les expériences menées dans les laboratoires cités ci-dessus sont sans exemples ailleurs.

Les premiers résultats dépassent les attentes. Quand on observe et qu'on analyse les formes nouvelles recueillies automatiquement, on éprouve une jubilation analogue, toutes proportions gardées, à celle qu'ont dû éprouver nos prédécesseurs quand ils mettaient pour

[*] ERSS, UMR 5610, CNRS & Université Toulouse II.
[**] ATILF, UMR 7718, CNRS & Université Nancy II.

la première fois côte à côte des paradigmes de langues indo-européennes ou quand ils observaient la répartition des aires dans les atlas linguistiques naissants. Dans le domaine de la phonologie lexicale en particulier, on découvre peut-être actuellement plus de généralisations nouvelles en un an que naguère en cinq ou dix années. Il ne fait guère de doute que, si cet effort de collecte et d'analyse était systématisé, notre connaissance des procédés de construction du français serait considérablement affinée, sinon renouvelée. Au-delà de ce renouvellement de l'empirie, de nouvelles questions voient le jour : l'accès aux formes les plus récentes permet notamment de mieux déterminer les contraintes actuellement à l'œuvre et de faire le départ entre les régularités figées du lexique attesté dans les dictionnaires et la morphologie vivante.

La taille du présent chapitre nous interdit de rendre compte de l'ensemble des recherches qui ont été menées pour automatiser la collecte des données et pour exploiter les résultats des collectes automatiques. Nous nous contenterons de décrire brièvement les outils utilisés et de donner une idée de la variété des trouvailles qu'ils permettent de faire. On montrera en particulier comment la constitution automatique de corpus – au sens technique du terme (*cf.* Habert *& alii* 1997) – a permis quant à elle de mettre au jour dans le domaine bio-médical des classes de formes qui échappent aux régularités à l'œuvre dans la langue générale.

2. Collecter des données sur la Toile

Traditionnellement, les morphologues travaillaient sur des ensembles d'attestations compilées à partir de dépouillements manuels d'œuvres lexicographiques ou littéraires. L'informatisation des dictionnaires et des textes rend aujourd'hui possible l'automatisation de cette collecte. Quelques secondes suffisent, par exemple, pour trouver l'ensemble des adjectifs en *-able* traités dans le *TLFi* (*Trésor de la Langue Française informatisé*), et il ne faut que quelques minutes pour extraire les formes en *-ette* d'une dizaine d'années d'archives du *Monde*. Il y a quelques années, ces tâches auraient demandé l'une plusieurs jours, l'autre plusieurs mois.

À côté de ces corpus traditionnels, les linguistes utilisent de plus en plus couramment la Toile pour trouver des attestations de lexèmes construits ou de tournures syntaxiques. L'exploration de la Toile peut se faire manuellement en interrogeant un moteur de recherche comme Google, Yahoo ou Exalead. Il s'agit alors de découvrir des occurrences de mots dont on prédit la possibilité en s'appuyant sur l'intuition. Il n'est pas rare de trouver ainsi nombre de formes pourtant réputées impossibles pour des raisons théoriques. Ainsi, par exemple, trouve-t-on préfixés par *anti-* même des adjectifs simples (*anti-obèse*, *anti-triste*) et des adjectifs en *–able* ou *–ible* (*anti-inflammable*, *anti-explosible*).

La collecte d'attestations sur le Web peut également être effectuée au moyen d'outils. Un projet comme WebCorp (http://www.webcorp.org.uk) met la technologie de base d'un concordancier à l'échelle du Web, en se fiant aux moteurs de recherche génériques. Ces moteurs sont actuellement le seul véritable moyen d'accès aux pages. Des systèmes de parcours et d'indexation spécifiques sont en cours de développement, mais ils ne pourront de toute façon pas prétendre à l'exhaustivité de Google ou de Yahoo. Nous avons, quant à nous choisi de développer des outils dédiés spécifiquement à la recherche lexicale : WaliM (Namer 2002) et Webaffix (Tanguy & Hathout 2002 ; Hathout & Tanguy 2002). Ces deux outils n'échappent pas à la règle et opèrent comme tous les autres systèmes sur la partie de la Toile – dont la proportion est d'ailleurs inconnue – à laquelle donnent accès les moteurs de recherche génériques, et donc sur un sous-ensemble de la Toile variable avec le temps, et sans critère de sélection identifiable. WaliM et Webaffix sont très proches par leurs objectifs et par leur fonctionnement général. Tous les deux génèrent des requêtes par programme, les soumettent à un moteur de recherche, puis effectuent différents nettoyages sur les résultats ramenés.

2.1. La boîte à outils Webaffix

Webaffix est à la fois une méthode et une boîte à outils d'enrichissement lexical à partir de la Toile. L'ensemble est destiné à la création et à la complétion semi-automatique de lexiques au moyen de collectes de formes construites par la morphologie. Il propose deux grandes fonctionnalités : la recherche de ces formes sur la Toile et leur filtrage, c'est-à-dire l'élimination d'une partie des réponses erronées ramenées par le moteur de recherche. La boîte à outils contient trois composants :

1. un module de recherche par suffixe qui permet de découvrir sur le Web des formes qui correspondent à un motif tel que la présence d'un suffixe graphémique donné[1] ;
2. un composant de prédiction morphologique permettant de calculer les formes des lexèmes bases ou des lexèmes construits ;
3. un méta-moteur disposant de fonctionnalités dédiées à l'exploration lexicale du Web.

Webaffix propose deux modes de recherche de formes nouvelles. Le premier exploite un lexique existant ou une base de données morphologiques, à partir desquels est générée une liste de formes candidates ; puis il utilise le méta-moteur pour rechercher sur la Toile des attestations de ces formes. Le second recourt au module de recherche par suffixe pour repérer sur le Web des formes qui correspondent à un schéma donné (par exemple celles qui

[1] Ce module faisait appel au moteur AltaVista, qui suite à son rachat par un concurrent, ne permet plus l'utilisation des caractères jokers nécessaires à la recherche par suffixe.

finissent par *-able*), sans aucune contrainte sur la base. C'est de cette façon, par exemple, que Hathout & *al.* (2003) ont découvert quantité de formes en *-able* dont le radical n'apparaît ni parmi les verbes ni parmi les noms du *TLF* et du *GRLF.*

Outre les problèmes classiques en détection de formes nouvelles collectées au moyen d'une recherche par suffixe (noms propres, xénismes, fautes d'orthographe, mécoupures, etc.), se pose la question du statut morphologique des résultats : s'agit-il bien de formes construites par le procédé morphologique auquel on s'intéresse ? Pour éliminer les formes qui ne sont pas dérivées, Webaffix propose deux méthodes de filtrage reposant sur l'existence ou l'inexistence des lexèmes-bases des mots collectés. La plus simple consiste à prédire les formes que prennent ces lexèmes-bases et à en rechercher des attestations sur la Toile. Dans ce cas, le critère pour le filtrage d'une forme comme e.g. *copolymérisable* sera l'attestation sur la Toile d'une des formes du verbe *copolymériser*. Un filtrage plus strict est également disponible : rechercher des pages Web qui contiennent à la fois la forme candidate et l'une des formes de son lexème-base.

2.2. Le système WaliM (Web et validation en Morphologie)

WaliM est un système plus simple qui fait appel au moteur de recherche Yahoo pour trouver des attestations des mots-requêtes qui lui sont fournis en entrée. En sortie, ces candidats sont répartis en trois groupes : ceux qui ne sont attestés nulle part sont potentiellement des mots construits « impossibles » ; ceux qui apparaissent dans un nombre suffisant de pages Web sont enregistrés comme corrects ; ceux pour lesquels le nombre des attestations est inférieur à un seuil donné sont considérés comme moins fiables. Pour ces derniers, les résultats sont filtrés en supprimant les caractères non alphanumériques dans les mots de la première page ramenée, puis en vérifiant que le mot-requête est bien toujours présent dans cette page.

2.3. Le Web comme corpus

Le corpus sur lequel opèrent WaliM et Webaffix est la Toile, et non un corpus au sens où l'entend la « linguistique de corpus ». Comme le note G. Grefenstette (1999), nombre de linguistes peuvent, à juste titre, se montrer réticents dans l'emploi de la Toile comme source d'attestations, étant donné en particulier l'impossibilité technique de caractériser automatiquement les pages du point de vue du domaine, du genre, du statut de l'auteur, ou de la validité du contenu.

Il convient de souligner aussi que les études quantitatives sur la Toile sont sujettes à caution. Le manque de contrôle sur les documents et l'absence de caractérisation globale de ce pseudo-corpus posent notamment des limites à la simple notion de fréquence. Les moteurs de recherche indiquent bien le nombre des documents retournés en réponse à une requête donnée, mais ce nombre n'est qu'une approximation grossière de la fréquence du mot qui constitue la requête. Il y a à cela plusieurs raisons. Tout d'abord, l'unité du Web

indexé est le document, et non pas l'occurrence lexicale : le nombre de documents gomme donc les répétitions d'une occurrence au sein d'un même document, et la notion d'hapax est elle-même rendue floue. Ensuite, le Web est le lieu par excellence de la reprise, de la citation et du plagiat. Des documents strictement identiques, mais accessibles à des adresses différentes seront considérés comme autant de réponses, augmentant artificiellement la fréquence d'un terme qui s'y trouverait. Le même phénomène s'applique à des citations partielles ou totales de documents. Des expressions comme *Au commencement était le verbe* ou *Mignonne, allons voir* apparaissent plusieurs dizaines de milliers de fois chacune. Enfin, il faut ajouter à ceci l'opacité des méthodes de calcul des moteurs de recherche, voire dans certains cas l'augmentation artificielle des nombres indiqués pour des raisons de marketing (voir à ce sujet les différents articles de J. Véronis sur son blog « Technologies du Langage »[2]. De ce fait, la comparaison des fréquences d'unités lexicales sur le Web est à manier avec précautions. Si la distinction présence/absence (*i.e.* aucun document *vs. n* documents renvoyés) peut être prise en compte, tout comme les différences en ordre de grandeur (2 documents *vs.* 100 000 documents), les différences faiblement marquées ne sont *a priori* pas significatives.

Ces mises en garde nécessaires n'interdisent nullement d'utiliser la Toile avec profit. Dans beaucoup de recherches, le nombre des attestations n'intervient pas ou, s'il intervient, reste maîtrisable. La morphologie constructionnelle fait grand usage de listes de formes ou de listes de constructions en faisant abstraction du nombre des échantillons qui représentent chaque type ; il est le plus souvent aisé de vérifier la fiabilité de chacun de ces types, même lorsqu'ils sont plusieurs milliers ; on aboutit ainsi à des listes dix, vingt ou trente fois plus longues que les listes compilées à l'aide des dictionnaires. Dans ces listes sont représentés de très nombreux cas de figure qui n'apparaissent jamais dans les données ordinaires et qui en apprennent long sur les contraintes phonologiques et sémantiques auxquelles est soumise la création lexicale. Dans d'autres cas, la simple attestation de quelques formes ou de quelques dizaines de formes suffit à démontrer la vraisemblance ou le bien fondé de telle ou telle prédiction. A quoi s'ajoute le fait que le nombre des textes auxquels la Toile donne accès s'accroît constamment et rapidement ; cette augmentation du nombre des données permet de revenir périodiquement sur certaines questions et de corriger ou d'affiner les hypothèses que suggérait une première description.

Ce serait se leurrer que de considérer le Web comme un corpus de « langue générale ». Si la variété des domaines abordés, et donc des sous-langages de spécialité représentés peut paraître suffisante, nous n'avons pas d'idée claire de la représentation de chacun de ces domaines. Toutefois, le simple fait de pouvoir constater que la langue varie et de pouvoir déterminer dans quel espace s'inscrit cette variation constitue souvent déjà un

[2] Notamment les articles « Les comptes bidons de Google » et « Moteurs : folles duplications » sur le site : http://aixtal.blogspot.com/.

progrès pour la morphologie constructionnelle. Et rien n'interdit de constituer de véritables corpus dans tel ou tel domaine particulier de façon à dégager les spécificités des types que ce domaine privilégie.

3. Vers une morphologie extensive

Quand, il y a une dizaine ou une douzaine d'années, est apparue la possibilité de réunir commodément de grandes quantités de formes absentes des dictionnaires, rien n'assurait que ce brusque afflux de données nouvelles amènerait des progrès dans notre connaissance de la morphologie du français : ces données auraient pu ne comporter que peu de faits nouveaux, ou les faits nouveaux se révéler ininterprétables. La langue est très répétitive et les formes isolées induisent bien souvent en erreur. Mais il est maintenant hors de doute que les efforts consentis pour mécaniser les récoltes dans les territoires immenses des textes électroniques ne sont pas vains. Les recherches récentes montrent que cette approche « extensive » permet d'établir des généralisations nouvelles de plus en plus fines au fur et à mesure qu'augmente le nombre des formes engrangées. On peut aussi, grâce à elle, documenter des faits rares dont l'existence demeurait incertaine du fait de la faiblesse de nos intuitions. Enfin, il devient possible de concevoir des dispositifs expérimentaux susceptibles de renouveler au moins en partie l'assise empirique de la discipline.

3.1. Généralisations nouvelles

C'est dans le domaine de la morphophonologie que les progrès ont été les plus rapides. Les résultats les plus spectaculaires sont sans doute ceux qui concernent les phénomènes de dissimilation. À titre d'exemple, nous relatons ci-après la façon dont s'est peu à peu dévoilé le conditionnement de la chute de certaines rimes devant le suffixe *–esque*. Dans le domaine de la sémantique, les progrès sont plus lents, mais quelques exemples, comme celui de la dérivation en *–able*, montrent qu'il y a lieu de s'attendre à ce que certaines questions puissent être renouvelées.

Les voyelles moyennes devant *–esque*. Les données qui ont permis le plus grand nombre d'observations nouvelles sont sans doute celles qui sont rassemblées dans la base de dérivés en *–esque* constituée à l'ERSS, en partie à l'aide de Webaffix, sous la responsabilité de Nicole Serna. Cette base réunit actuellement environ 3 000 formes distinctes accompagnées chacune d'un ou plusieurs exemples référencés. (À titre de comparaison, le *Robert électronique* et le *TLFi* contiennent l'un et l'autre moins d'une centaine de tels dérivés). Les enseignements que fournit cette base sur la morphophonologie du français sont multiples. Pour illustrer notre propos, nous nous contenterons de relater ici la façon dont est apparu puis s'est précisé peu à peu un problème tout à fait inédit : celui du comportement des finales en voyelle antérieure moyenne (/e, ɛ, ø ou œ/) + consonne fixe devant *-esque*.

À ne se fier qu'aux dictionnaires, les bases se terminant par une finale de ce type ne soulèvent aucun problème particulier, puisque les quelques dérivés attestés dans les ouvrages cités ci-dessus (*babélesque*, *moliéresque*, *raphaélesque*) sont formés par simple concaténation du suffixe au lexème-base tel qu'il apparaît à l'état libre.

L'idée que l'identité (ou la similitude) de la voyelle suffixale et de la voyelle finale du lexème-base puisse entraîner à certaines conditions la chute de cette dernière n'est apparue que vers 1995, à un moment où la base de données comprenait environ 800 dérivés. Cette nouvelle liste suggérait en effet que les rimes en /ɛ/ + consonne fixe peuvent parfois disparaître, à partir du moment où la base fait quatre syllabes, comme dans *nibelungesque* et *pantagruesque* (*cf.* Plénat 1997 : 168). On pouvait constater que les bases tétrasyllabiques en -*eur* étaient elles aussi accourcies (cf. *consommatesque*, *déprédatesque*), mais, curieusement, les données comportaient un cas où la finale –*eur* disparaissait à la fin d'une base de seulement trois syllabes : *tirailleur* → *tiraillesque*. Le trissyllabe *Cervantes* était également accourci dans *cervantesque*, mais ce comportement ne se distinguait pas du comportement le plus courant des trissyllabes s'achevant par /s/ (cf. *clitoresque*, *cosinesque*). Quelques années plus tard, la récolte d'environ 400 formes nouvelles n'avait pas permis de progresser : Plénat (2000 : 32) constate la relative labilité des finales en voyelle antérieure moyenne + consonne fixe devant –*esque* à la fin des bases longues, mais sans pouvoir déterminer précisément le seuil à partir duquel l'accourcissement a lieu.

La base de données actuelle permet de donner des faits une description sensiblement plus précise :

1. Il se confirme que les rimes en voyelle d'avant moyenne + consonne fixe peuvent tomber devant –*esque* quand le lexème-base comprend quatre syllabes : on a à la fois, par exemple, *polichinellesque* et *polichinesque*, *harrypotteresque* et *harrypottesque*, *vétérinaresque* et *vétérinesque*, *ordinatoresque* et *ordinatesque*. Il se confirme également qu'en règle générale, ces rimes se maintiennent lorsque la base est plus courte. Néanmoins,
2. Elles peuvent tomber aussi dès que la base fait trois syllabes non seulement lorsque la dernière consonne est identique à l'une des consonnes suffixales (cas de *cervantesque* ou de *BTesque*, formé sur *BTS*), mais aussi si cette consonne est déjà représentée au moins une fois dans la base : par exemple, *Ben Laden* (2 /n/) donne *benladesque* (et *benladénesque*), *colonel* (2 /l/) donne *colonesque*, *Internet* (2 /t/) donne *internesque* (et *internetesque*), *Warhammer* (2 /r/) donne *warhammesque* (et *warhammeresque*). Autrement dit, l'énigme de la dérivation *tirailleur* → *tiraillesque* est résolue : c'est la présence de deux /r/ dans le lexème-base qui entraîne la chute de -*eur*.
3. Enfin, elles tombent même si la base ne compte que deux syllabes lorsque la consonne fixe finale est identique (ou quasi-identique) à l'une des consonnes du suffixe : avec les sifflantes, on a *(Fabien) Barthez* → *barthesque*, *(Edmond) Dantès* → *dantesque*,

(Louis de) Funès → *funesque* ; avec /k/, on a *Cherek* (une île imaginaire) → *cheresque*. Il arrive que d'autres rimes en sifflante disparaissent à la fin de bases dissyllabiques (comme dans *phidiesque* ou *pouffiesque*), mais ces disparitions sont, semble-t-il, moins systématiques que lorsque la dernière voyelle de la base est un /ɛ/ ; et, ordinairement, les rimes en /k/ ne commencent à tomber que lorsque la base fait trois syllabes (cf. *goldoresque*, *mobydesque*).

Comme on le voit, l'accroissement de la quantité de données disponibles a un rôle analogue à celui de l'introduction du microscope dans les sciences de la nature. Là où l'inspection du petit nombre de formes enregistrées dans les dictionnaires ne permettait pas de voir quoi que ce soit, un grossissement de 30× révèle une quantité plus qu'appréciable de faits nouveaux. Et ces faits font sens. Aux forces conservatrices qui tendent à préserver l'intégrité du lexème-base et du suffixe et qui imposent le plus souvent une simple concaténation des deux éléments, s'opposent deux sortes de contraintes : des contraintes de taille, qui pénalisent les formes de plus trois syllabes (*cf.* Plénat, ce volume) et des contraintes dissimilatives qui pénalisent la consécution dans une même forme de deux voyelles ou de deux consonnes identiques ou similaires (*cf.* Lignon & Plénat, ce volume). Aucune de ces contraintes ne contrebalance à elle seule les forces conservatrices, mais quand elles agissent de concert, elles peuvent aboutir à des troncations.

La plasticité sémantique des dérivés en *–able*. L'augmentation de la quantité de données soumises à observation est également déterminante quand on s'intéresse aux dimensions catégorielle et sémantique de la description morphologique. Sur ce point, l'étude de la dérivation en *-able* de Hathout *& alii* (2003) donne d'assez bonnes indications sur le type de progrès que l'on peut attendre de l'accumulation de données nouvelles.

Les études antérieures de la dérivation en *-able* (Leeman & Meleuc 1990, Leeman 1992, Anscombre & Leeman 1994, en particulier) se fondaient principalement sur des relevés lexicographiques et sur un recours massif aux jugements de grammaticalité. Le corpus disponible était d'environ 1 400 formes, ce qui correspond en gros au nombre d'adjectifs en *-able* présents dans la nomenclature d'un grand dictionnaire comme le *GLLF*, le *GRLF* ou le *TLF*. Hathout *& alii* (*art. cit.*) ont constitué pour cette étude une liste de près de 5 000 adjectifs en utilisant successivement les deux méthodes de collecte de Webaffix et, pour certains dérivés, ont analysé systématiquement les emplois présents sur la Toile.

Les dérivés en *-able* ont souvent été considérés comme des déverbaux de « sens passif » (autrement dit, leur nom recteur était dit correspondre à un objet direct du verbe de base ou à un patient, selon que l'on considérait la syntaxe ou la structure argumentale). Nos récoltes montrent que, si tel est bien la plupart du temps le cas, le nom recteur peut en fait représenter un grand nombre d'autres types de participants au procès. Le plus simple, pour illustrer cette plasticité, consiste sans doute à passer en revue les noms recteurs d'un dérivé

comme *pêchable* (qui n'apparaît pas dans le *TLF*). Il va de soi que sont pêchables en premier lieu les poissons, mollusques ou batraciens qui peuplent les eaux. Mais le sont aussi certains lieux : ceux où peuvent se poster les pêcheurs (berges, ponts, digues, etc.) et ceux où circulent leurs proies (rivières, étangs, courants, etc.). Selon que la pêche est ouverte ou non, suivant qu'il fait beau ou mauvais, les saisons, les jours et les circonstances atmosphériques peuvent eux aussi être ou non pêchables. On trouve également des contextes où *pêchable* est prédiqué du matériel de pêche (mouches ou nylon, par exemple). Enfin, ce ne sont pas seulement les participants du procès, mais aussi les propriétés de ces participants qui peuvent être qualifiées de pêchables ou d'impêchables : nous avons des exemples où c'est la taille de certains poissons qui est dite impêchable, et nous ne désespérons pas de trouver un jour des *courants d'une violence impêchable* ou *des cannes d'une rigidité difficilement pêchable*. Il n'y a guère que les pêcheurs qui, en tant qu'agents, ne peuvent apparemment pas être dits pêchables, et la dérivation en *–able* apparaît ainsi comme l'inverse de la dérivation en *–eur*, laquelle ne fournit au contraire que des agents (Fradin & Kerleroux, ce volume).

Quand on constate l'existence, la prolifération même, dans la langue commune, de tels emplois, qui n'avaient jamais été répertoriés, certains emplois « circonstanciels » mieux connus deviennent plus intelligibles : on comprend mieux que les *verba aedificandi* ou *movendi* donnent des adjectifs en *-able* dont le nom recteur désigne un lieu (cf. *un terrain constructible*, *bâtissable*, *une piste skiable*, *roulable*), les *verba laborandi* des dérivés dont le nom recteur désigne une période (cf. *des jours ouvrables*, *travaillables*), les *verba damnandi* des adjectifs dont le nom recteur désigne le chef d'inculpation (cf. *un tour pendable, une folie enfermable*), ou encore qu'une propriété comme le prix puisse être dite abordable (cf. *une jupe d'un prix abordable*).

Les dérivés en *–able* sur base substantivale. Les descriptions classiques signalent l'existence d'un certain nombre d'adjectifs en *-able* construits sur des bases substantivales. Ainsi Gawelko (1977) signale, outre certains dérivés de noms abstraits – noms de qualité (cf. *charitable*) ou noms d'action (cf. *viable*) –, l'existence de petites séries dérivant de noms d'impôts (*corvéable*, *mainmortable*), de véhicules (*carrossable*, *cyclable*) ou de titres (*consulable*, *papable*). Notre enquête nous confirme l'existence de ces séries : les impôts modernes (cf. *TVable*, *ISFable*), les véhicules récents (cf. *jeepable*, *planchable*) et les titres de toutes sortes (cf. *rectorable*, *chairable*, *étoilable* (de (*danseuse*) *étoile*)) fournissent en abondance des dérivés en *-able*. Mais il y a plus : d'autres adjectifs dénominaux, apparaissent qui concordent avec des séries de dérivés déverbaux remarquables ou moins remarquables. De la même façon que *carrossable* ou *jeepable* peuvent être rapprochés de dérivés de *verba movendi* comme *skiable* ou *roulable*, de la même façon *piscinable* ou

boxable (dans *un terrain piscinable*, *un garage boxable*) doivent être mis en parallèle avec des dérivés de *verba aedificandi* comme *constructible*, et *peinable de mort* (dans *un crime peinable de mort*) avec des dérivés de *verba damnandi* comme *pendable*. D'autres dérivés dénominaux sont remarquables moins par les noms recteurs qu'ils qualifient que par la fonction thématique de leur base, qui, par exemple, peut être un lieu (cf. *une statue muséable*, *un ministre matignonnable*) ou un état final (cf. *des pommes de terre fritables*, *un lait fromageable*). En un mot, l'enquête confirme avec éclat que les contraintes catégorielles ont une origine sémantique : les dérivés en *-able* sélectionnent habituellement des bases verbales parce que celles-ci dénotent des procès ; quand à un procès ne correspond aucun verbe, un nom fait l'affaire.

La constitution de bases de données comme celles que nous venons de mentionner prend un temps considérable. Si la collecte est désormais très rapide, la validation des données collectées exige un long travail philologique. Mais la preuve est faite que le pari est réussi : des généralisations nouvelles sont mises au jour non seulement dans le domaine de la phonologie, mais aussi dans celui des contraintes catégorielles et de l'interprétation sémantique.

3.2. Faits rares

Il n'y a pas lieu, à notre sens, d'opposer la recherche systématique d'attestations de formes nouvelles et l'intuition. Notre expérience nous suggère que les locuteurs sont capables de jugements très sûrs, même s'agissant de configurations trop rares pour qu'on ait une chance raisonnable de les rencontrer au cours d'une vie entière de lectures. Dans ce type de situation, cependant, les recherches sur la Toile permettent de conférer le statut de fait vérifiable à ce qui, autrement, serait resté en l'état de simple conjecture, car, grâce aux moteurs de recherche, on peut explorer en un temps très court des quantités de texte que des dizaines de vies ne permettraient pas de parcourir.

L'exemple sans doute le plus spectaculaire de ce genre de confirmation concerne la substitution de la finale *–este* au suffixe *–esque*. Analysant un exemple isolé de Verlaine (*silviopelliqueste*, de *Silvio Pellico*), Pichon (1940) dit son sentiment que cette substitution doit être comprise comme un phénomène de dissimilation prenant effet après une consonne vélaire. Au cours du demi-siècle suivant, cette conjecture a été citée plusieurs fois, mais sans qu'on puisse avancer un seul autre exemple de la substitution de *–este* à *–esque* après /k/ ou /g/. Récemment, toutefois, Plénat & *alii* (2002), grâce en partie à une recherche sur la Toile menée à l'aide de Webaffix, ont déniché une demi-douzaine d'exemples nouveaux. À l'heure actuelle, notre base de données contient une bonne trentaine de formes en *–gueste*

et -*queste*, dont certaines sont bien ou même très bien attestées (*titaniqueste*, *(jack)langueste*, ou *blogueste*, par exemple).

Pour illustrer cette possibilité que fournit la Toile de confirmer des intuitions incertaines, on nous permettra ici d'évoquer un point de détail de la morphologie des adverbes en *–ment*. On sait que ces adverbes adoptent en général pour thème celui du féminin de l'adjectif correspondant (cf. e.g. *fraîchement*, *nouvellement*, *rageusement*). On sait aussi qu'à l'ordinaire, les adjectifs en *–ant* et *–ent* donnent par exception des adverbes en -*amment* et -*emment* (cf. *méchamment*, *intelligemment*). Historiquement, cette particularité s'explique par le fait que, dans l'ancienne langue, la plupart de ces adjectifs n'avaient qu'une forme pour les deux genres. Il n'est cependant pas vrai que tous les adjectifs en *–ant* et -*ent* donnent des adverbes en -*amment* et en -*emment*. Ainsi, s'appuyant sur une remarque de Molinier (1992), Yvon (1996 : 164) note que *charmamment* et *clémemment* sont d'une acceptabilité pour le moins douteuses. Nous avions, quant à nous, le sentiment que ces deux adverbes sont purement et simplement agrammaticaux et que, pour adverbialiser *charmant* et *clément*, il convient de recourir, comme dans le cas général, au thème de féminin et de dire *charmantement* et *clémentement*. Cette agrammaticalité des formes suggérées par les grammaires s'explique, pensons-nous, par le fait qu'elles contiennent deux /m/ et deux voyelles analogues (/mamã)/) dans deux syllabes consécutives : c'est une contrainte dissimilative qui impose le choix d'un thème inattendu avec *–ant* et *–ent*.

Nous avions jusqu'à présent renoncé à faire état de cette hypothèse, faute de pouvoir apporter d'argument véritablement convaincant. Le seul dont nous disposions résidait en la constatation que, parmi les trois exceptions citées par les grammaires, figure *véhémentement*. Une récente interrogation de la Toile à l'aide de Webaffix apporte maintenant de l'eau à notre moulin. *Charmamment* n'est attesté qu'une fois, dans une liste de discussion où le locuteur reconnaît peu après qu'il aurait dû écrire *charmantement*. Cette dernière forme, en revanche, apparaît quant à elle dans une dizaine de bons exemples. En outre, une recherche rapide dans *Frantext* nous a appris qu'Albert Cohen a lui aussi utilisé *charmantement*, dans son roman *Mangeclous*. Pour ce qui est des dérivés de *clément*, nous avons trouvé un excellent exemple de *clémentement* ; nous avons également trouvé une fois *clémemment*, mais dans un dictionnaire français-anglais qui applique aveuglément la règle des grammaires. Nous avons en outre découvert une très bonne attestation de *aimantement* et une demi-douzaine d'exemples convaincants de *démentement*. Pour être complets, nous avons aussi trouvé un i*nfamamment* et une demi-douzaine de *alarmamment*, mais dans de mauvaises traductions, nous semble-t-il. Ces quelques remarques n'épuisent certainement pas la question, mais tout nous porte à croire désormais que notre intuition ne nous trompait pas et que le français recourt à des adverbes en *–mantement* et *–mentement* plutôt qu'en *–mamment* et -*memment*, quand du moins on ose recourir à des néologismes.

La discussion précédente montre aussi que, de toute façon, le recours à des jugements d'acceptabilité reste indispensable : les textes électroniques sont de qualité trop inégale pour que l'on accueille sans examen toutes les attestations.

3.3. Vers une morphologie expérimentale

Les observations qui fondent les travaux évoqués ci-dessus ou les expériences du même genre peuvent être reproduites, soit sur les mêmes données si celles-ci ont été mises à la disposition du public, soit sur des données comparables dans la mesure où leur mode d'obtention a été décrit. Et les prédictions qu'autorisent ces analyses peuvent être testées systématiquement sur des données nouvelles. Le lecteur peut, par exemple, obtenir quasi instantanément des attestations de *benladesque*, de *blogueste* ou de *démentement* en soumettant ces requêtes à un moteur de recherche, établir la liste des adjectifs en *–mant* et en *–ment* et rechercher les adverbes correspondants, ou bien encore déterminer quels noms recteurs apparaissent devant *chassable* ou *navigable* et voir si leurs types sont aussi variés que ceux des noms recteurs de *pêchable*.

Mais prenons un exemple supplémentaire. Dal & Namer (2005) ont récemment monté une expérience sur une sorte d'« échangisme entre bases » observable devant le suffixe *–ité*. Leur point de départ réside dans le constat que, bien que ce suffixe exige à l'ordinaire une base adjectivale, on trouve assez souvent comme bases des toponymes au lieu des adjectifs correspondants : *ivoirité*, par exemple, est beaucoup plus fréquent que *ivoirianité*, et *portugalité* supplante entièrement, semble-t-il, *portugaisité*. Afin de déterminer la répartition de ces deux types de bases, les autrices ont réuni une centaine de noms de pays et de régions avec leurs allomorphes (e.g. *Chine~sin-* ; *Danemark~dan-*) ainsi que les adjectifs ethniques correspondants, dérivés ou non (e.g. *Hongrie*, *hongrois*, *magyar*). Elles ont ensuite construit sur ces formes tous les dérivés en *-ité* phonologiquement vraisemblables, puis, à l'aide de WaliM, déterminé et quantifié la présence sur le Web de chacune de ces formes-candidates.

Les résultats, frappants, opposent deux classes de suffixes : quand l'adjectif ethnique comporte les suffixes *-ain*, *-ien* ou *-éen*, c'est lui, dans sa variante « savante » qui, en général, est retenu comme base du dérivé en *-ité*, de préférence au nom de pays (e.g. *marocan-ité*, *italian-ité*, *coréan-ité*) ; quand, en revanche, le gentilé comporte les suffixes *-ois* ou *-ais*, c'est ou bien un adjectif supplétif (*magyar-ité*) ou bien le toponyme (ou encore une variante liée de celui-ci) qui sert systématiquement de base (*franc-ité*, *dan-ité*, *sin-ité*). Dans les autres cas de figure, les autrices ne discernent aucune régularité.

Malgré ce qui vient d'être dit, il existe une classe d'adjectifs ethniques en *-ien* (ou *-éen*) qui, selon Dal & Namer (*ibid.*) ne pourraient pas servir, du moins tels quels, de bases à une suffixation en *-ité* : ceux dans lesquels le suffixe est précédé de /n/ : *Lusitanie* donne *lusitanité* et non *lusitanianité*, et l'on a de même, selon les autrices, *estonité*, *iranité*,

jordanité, *mauritanité*, *méditerranéité*, *palestinité*, *ukrainité*. On a sans doute affaire là à un phénomène de dissimilation préventive, puisque le choix du toponyme comme base permet d'éviter la consécution de deux /n/ dans le dérivé.

Ce cas de dissimilation préventive est particulièrement intéressant. Comme dans *tiraillesque* ou *benladesque*, c'est la crainte d'une répétition interne à la base qui provoque le phénomène. Nous avons donc voulu, quelques mois après l'expérience initiale, vérifier la stabilité de ce résultat sur la Toile. Tous les exemples fournis par Dal & Namer (*ibid.*) ont été confirmés (à ceci près que nous avons rencontré quelques attestations de *palestinianité*). Nous avons en outre trouvé *arménité*, *étasunité*, *ghanéité*, *guinéité*, *macédonité*, au lieu de *arménianité*, *étasunianité*, etc. Enfin, pour déterminer si la crainte d'une répétition de /n/ pouvait aboutir à la disparition du suffixe adjectival devant d'autres suffixes nominaux, nous avons mené une recherche du même type avec le suffixe *-itude*, et nous avons trouvé *arménitude*, *bosnitude*, *calédonitude*, *californitude*, *estonitude*, *macédonitude*, *palestinitude*, *ukrainitude* (et même *bourguignitude* !), sans jamais rencontrer les *arménianitude*, *bosnianitude*, ou *bourguignonitude* attendus. L'avenir permettra peut-être d'affiner ce résultat, mais on peut, pensons-nous, le considérer comme acquis.

Le recours à l'intuition est lui aussi une sorte d'expérience. Mais c'est, en quelque sorte, une expérience privée et souvent incertaine, qui n'est pas remise en cause par des intuitions ultérieures ou extérieures. La répétibilité et la netteté d'expériences comme celle qui vient d'être décrite emportent la conviction. On peut préférer *palestinianité* à *palestinité*, mais on ne peut mettre en doute l'existence de la dissimilation chez un grand nombre de locuteurs.

4. Sémantique des composés néoclassiques

Les différentes études présentées ci-dessus portent toutes sur la morphologie dans la « langue générale ». Aucune restriction particulière n'est imposée sur les domaines ou les genres des documents dans lesquels les exemples sont collectés. Cette absence de contraintes sur les sources facilite la constitution de bases morphologiques de grande taille comme celles des dérivés en *-esque* ou en *-able*. Mais l'augmentation du nombre des données est tout aussi fructueuse dans des corpus mieux contrôlés. Les travaux de F. Namer sur la composition dite néoclassique illustre parfaitement cette approche. En étudiant des corpus spécialisés de grande taille, elle a mis en lumière des propriétés sémantiques et structurelles de lexèmes construits par cette composition que les études théoriques antérieures n'ont pas pu découvrir, faute d'accès à ces données.

Ces corpus ont été réunis dans le cadre du projet UMLF[3], à partir de documents électroniques de toutes sortes : littérature scientifique, comptes-rendus hospitaliers anonymisés, lexiques, thésaurus, etc. Ils totalisent environ 12 millions d'occurrences. Leur vocabulaire comporte quelque 209 000 lexèmes, dont une grande partie est morphologiquement complexe.

4.1. Arrière-plan théorique

Qu'ils soient « populaires » ou « néoclassiques », les composés nominaux se laissent en règle générale décrire sémantiquement en fonction de deux types de rapports : le rapport qu'entretient le composé avec ses composants et le rapport qu'entretiennent les composants entre eux. Suivant que le composé hérite ou non de ses composants, ou de l'un d'entre eux, son type sémantique, on peut parler de composés « primaires » ou « secondaires ». Suivant que les composants sont sur un pied d'égalité ou, au contraire, que l'un est dans la dépendance de l'autre, on parle de composés « copulatifs » ou « déterminatifs »[4].

		Populaires	Néoclassiques
Primaires	copulatifs	*moissonneuse-batteuse*	*rhinopharynx*
	déterminatifs	*poisson-chat*	*métropole*
Secondaires		*allume-cigare* (*casque-bleu*)	*sauroctone* *pachyderme*

Tableau 1. Types de composés

On parle de type copulatif quand le sens du composé est la conjonction des sens de ses parties (une moissonneuse-batteuse est une machine qui moissonne et qui bat les céréales, *rhino-pharynx* désigne l'ensemble constitué par le nez (°*rhino*-) et le pharynx). Quand le composé n'est pas de type copulatif, l'un des composants est déterminé par l'autre. Cette détermination peut prendre plusieurs formes : si l'élément déterminé dénote un procès, le déterminant dénote un des participants de ce procès (*allume-cigare* désigne un dispositif qui permet d'allumer pipes, cigares et cigarettes ; l'épithète *sauroctone* est appliquée à Apollon tueur (°-*ktone*) de lézards (°*sauro*-) ; quand l'élément déterminé est un objet, le déterminant dénote une propriété constitutive ou fonctionnelle qui caractérise cet objet (un poisson-chat est un poisson qui a l'aspect d'un chat, un pachyderme est un animal dont la peau (°*derme*) est épaisse (°*pachy*-). Un composé copulatif relève de la même classe sémantique que ses composants (par exemple, le rhinopharynx est une cavité anatomique comme le nez et le

[3] Le projet ACI UMLF « Lexique médical francophone unifié », a été financé par le MNERT de 2002-2004, et piloté par P. Zweigenbaum (INSERM), (*cf.* Zweigenbaum & *al.* 2005).

[4] Nous reprenons ici des distinctions traditionnelles qui s'inspirent de celles que faisaient les grammairiens indiens (*cf.* Renou 1930 : 82 *sqq.*).

pharynx). Les composés déterminatifs se subdivisent quant à eux en deux catégories : les composés primaires endo-centriques, dont l'un des composants relève de la même classe sémantique que l'ensemble et entretient avec celui-ci une relation d'hyperonyme à hyponyme, et les composés secondaires exocentriques, dont les composants appartiennent l'un et l'autre à une classe différente de celle du composé. Par exemple, *poisson-chat*, qui désigne un poisson et *métropole*, qui désigne une ville (°*pole*) sont endo-centriques, alors que *pachyderme*, qui ne désigne pas un tégument mais un animal, ou *allume-cigare*, qui ne désigne pas un procès mais un dispositif, sont exocentriques.

Comme le montre le Tableau 1, la description ci-dessus s'applique tant aux composés dits « populaires » qu'aux composés dits « néo-classiques ». Ce qui distingue les seconds des premiers, c'est, d'une part, le fait qu'ils recourent en général à des lexèmes empruntés au latin ou au grec (au grec dans les exemples du Tableau 1), et, d'autre part, que, contrairement aux composés « populaires », ils se conforment ordinairement au schème déterminant-déterminé et non au schème déterminé-déterminant quand ils ne sont pas de type copulatif : cf. *métro**pole*** vs. ***poisson**-chat*, *sauro**ctone*** vs. ***allume**-cigare*, *pachy**derme*** vs ***casque**-bleu* (qui a des chances d'être un syntagme lexicalisé, *cf.* Fradin 2003 : 199 *sqq.*). Ce parallélisme invite à faire l'hypothèse que le calcul du sens des composés de l'une et l'autre classes obéit aux mêmes principes.

L'étude du sens des composés telle que la présentent par exemple Corbin (2004), Fradin (2000), Iacobini (2003), Villoing (2003) ou Warren (1990) se fonde sur l'idée que le sens des lexèmes construit est compositionnel. Plus précisément, un composé serait construit de façon binaire et tout composé comportant plus de deux composants relèverait d'une itération ou, parfois, d'une récursion des opérations de composition. Ainsi, *électro-cardiogramme* se définit par rapport à *cardiogramme* : c'est un tracé (des battements) du cœur (*cardiogramme*) obtenu par l'électricité (*électro-*), et un *mange-mange-sommeil* est un animal fantastique qui se nourrit (*mange-*) d'autres animaux fantastiques appelés *mange-sommeil*. De même, les interactions entre affixation et composition seraient analysables comme des successions d'opérations élémentaires dont l'ordre serait motivé linguistiquement. Par exemple *antivirus-thérapie* 'emploi thérapeutique des antivirus' conjoint le préfixé *antivirus* et *thérapie*, alors que *antipédophile* 'contre les pédophiles' est obtenu par préfixation de *anti-* au composé *pédophile.*

4.2. Données récalcitrantes du domaine bio-médical

Si les données propres au domaine médical obéissent en général aux hypothèses établies pour la langue générale, deux séries au moins de composés néoclassiques paraissent rétifs aux principes de compositionalité et de binarité énoncés ci-dessus. Ces données constituent des micro-séries (totalisant un millier de composés au plus) qui n'auraient pu être détectées sans la collecte puis l'analyse de données massives spécialisées. L'observation de ces

données « non conformes » conduit à se demander dans quelle mesure la composition néoclassique constitue un procédé ressortissant à la morphologie constructionnelle du français, d'autant plus que ces mêmes séries se rencontrent dans la plupart des langues européennes.

La première série comprend deux sortes de composés déterminatifs à trois composants, dans lesquels le dernier composant dénote une action médicale et entretient un rapport d'hyperonyme à hyponyme avec l'ensemble, qui dénote donc lui aussi une action médicale. Dans ces noms, les deux premiers composants dénotent des participants au procès. Dans le premier type, ces participants jouent le même rôle sémantique et peuvent être intervertis, comme par exemple dans les synonymes : *urétro-cystoplastie* et *cysto-urétroplastie* : « reconstruction chirurgicale (*plastie*) de la vessie (*cysto*) et de l'urètre (*urétro*) ». Ce premier type comprend de nombreux noms qui expriment des activités d'observation (*-graphie, -métrie*) ou des interventions chirurgicales (*-pexie, -rrhaphie*). Dans le second type, en revanche, les participants ne jouent pas le même rôle. Il s'agit le plus souvent de composés en *-stomie* où les deux premiers composants indiquent les points de départ et d'arrivée d'une ouverture pratiquée chirurgicalement (c'est là le sens de *stomie*) : ainsi une *duodénoentérostomie* permet de faire communiquer le duodénum avec une partie de l'intestin grêle (*entéro*). Le fait que *entéroduodénostomie* est également attesté, avec le même sens, montre que les deux premiers composants sont non-ordonnés et appartiennent au même niveau d'analyse.

Le second ensemble de termes paraissant obéir à des principes différents de ceux qui valent pour les composés de la langue générale comprend des noms à la fois composés et préfixés pouvant être notés linéairement par *pfx-Y-X-ie*. Dans ce schème, *pfx* représente l'un des trois préfixes quantificateurs *a-*, *hypo-* ou *hyper-*, et l'ensemble dénote une pathologie caractérisée suivant le préfixe par une absence, une insuffisance ou un excès de X et / ou de Y. La quantification porte en effet tantôt sur *X* (*agastrémie* veut dire 'absence de sang = ˚*ém* dans l'estomac = ˚*gastr*'), tantôt sur *Y* (*hypofibrinémie* signifie 'insuffisance de fibrine dans le sang = ˚*ém*'), tantôt encore sur la conjonction de *Y* et *X* (*acheiropodie* veut dire 'absence de mains = ˚*cheir* et de pieds = ˚*pod*'). Le sens du composé dépend non pas de l'ordre des composants, mais de la relation qu'entretiennent les référents de ceux-ci : si l'un est un constituant naturel de l'autre, la pathologie est due à la quantité anormale de ce constituant dans l'élément dans la constitution duquel il entre (*hyperprotéinémie* dénote un excès de protéines dans le sang, *anentérémie* une absence de sang dans l'intestin grêle). Quand, en revanche, les référents n'entrent pas dans la constitution l'un de l'autre, la pathologie est due à la quantité anormale des deux. Ce dernier cas d'ailleurs ne s'observe à notre connaissance qu'avec le préfixe *a-*, dans des composés qui dénotent des pathologies congénitales touchant l'embryon (*amyélencéphalie* renvoie à l'état caractérisé par l'absence de moelle épinière = ˚*myel* et de cerveau = ˚*encéphal*). Comme les

composés de la première série, ces composés comprennent donc deux composants non ordonnés qui relèvent du même niveau d'analyse. La langue générale ne possède pas de telles séquences de deux composants mis en relation avec un troisième sans que soit calculé le sens de l'unité qu'ils forment.

La mise en évidence des données que nous venons de décrire a été rendue possible, à partir de ressources textuelles spécialisées massives, grâce à un ensemble de techniques applicables en séquence, et grâce à l'organisation finale des données sous la forme d'une base de données (Namer 2005, Zweigenbaum *& al.* 2005). La consultation d'un bon dictionnaire de médecine aurait probablement permis d'aboutir à des constatations analogues, mais au prix de journées entières de lecture. Soulignons d'ailleurs qu'aucun dictionnaire médical, si complet fût-il, n'est en mesure de prédire l'ensemble des structures morphologiques du vocabulaire néoclassique, dont les éléments connaissent un renouvellement constant et naturel : seuls la numérisation des rapports cliniques ou des compte-rendus hospitaliers permet d'en conserver une trace. Il y a plus de dix ans déjà, des études (Lovis *& al.* 1995 ; Schulz *& al.* 1999) signalaient l'impuissance des dictionnaires spécialisés à prendre en compte la progression constante du vocabulaire morphologiquement construit, et en particulier par composition néoclassique, qui constituait alors, d'après Lovis *& al.* (1998), au moins 60% des néologismes présents dans les documents médicaux (rapports, littérature scientifique, etc.). Ces données font clairement apparaître que les hypothèses théoriques en matière de composition savante ne prennent pas en compte la réalité du vocabulaire biomédical. Celui-ci possède ses règles propres qui s'ajoutent à celles de la langue générale (et parfois les suppléent). Ces règles étaient sans doute déjà en vigueur en grec (où l'on disait βατραχομυομαχία (*batracho-myo-machie*) pour 'combat des grenouilles et des rats') et se sont probablement maintenues dans la langue des clercs (*cf.* la *Monachopornomachie* de Simon Lemnius), mais elles ne trouvent pas à s'appliquer, à notre connaissance du moins, dans les composés savants de la langue commune. Dans la mesure où le vocabulaire bio-médical relève d'un domaine de connaissance fortement structuré par les connaissances mêmes et la pratique de ses acteurs, ces règles se rapprochent de celles d'une terminologie. Ajoutons enfin que, selon Iacobini (2004), le même phénomène apparaît dans d'autres langues européennes, où s'observent des noms construits sur les mêmes modèles (cf. *uranostaphylorrhaphy*$_{EN}$, *Enterokolostomie*$_{DE}$, ou *iperproteinemia*$_{IT}$, *anenteroneuria*$_{ES}$)[5].

[5] Respectivement traduits par : *uranostaphylorraphie*, *entérocolostomie*, *hyperprotéinémie*, *anentéroneurie*.

5. Conclusion

Les progrès dans les sciences sont étroitement tributaires du contexte matériel et social dans lequel sont menées les recherches. L'introduction de l'informatique dans les laboratoires de linguistique ne pouvait rester sans effet. En morphologie, l'accès à de très grandes quantités de données textuelles, sur la Toile notamment, et l'utilisation d'outils capables de traiter ces données dans des délais raisonnables permettent de renouveler le socle empirique de la discipline. Il devient possible de multiplier des observations d'une finesse qui aurait été impensable il y a quelques années. On voit mal comment, sans l'assistance de l'informatique, on aurait pu, par exemple, déterminer la gamme des rôles thématiques des noms recteurs d'un adjectif comme *pêchable* ou construire une description dans laquelle la chute de la rime finale de *tirailleur* dans *tiraillesque* s'intègre naturellement. Dans les cas les plus favorables, ces observations convergent et laissent entrevoir le rôle que jouent un certain nombre de grands principes. Le lecteur aura par exemple noté l'importance que revêt en français le principe de dissimilation, qui intervient dans tous les exemples que l'on a donnés en morphophonologie (chute des rimes en voyelle moyenne devant *–esque*, dérivés adverbiaux des adjectifs en *-ment* et *–mant*, déverbaux en *–ité* et en *-itude* des gentilés en *–ien* et *-éen*). Si l'on multiplie les expériences comme celles qui ont été décrites, on peut espérer déterminer assez rapidement les grandes forces qui modèlent la création lexicale en synchronie, ce que ne permettaient certainement pas de faire les dictionnaires, qui sont encombrés par la compilation d'un héritage séculaire. Beaucoup de progrès restent à faire dans la collecte et le traitement des données. Ainsi conviendrait-il par exemple de mener de grandes enquêtes par domaines analogues à celles dont fait l'objet le domaine bio-médical. Mais on a le sentiment que la discipline vit dès maintenant les débuts d'une mutation importante.

Références

ANSCOMBRE, J.C., & D. LEEMAN (1994), « La dérivation des adjectifs en *-ble* : morphologie ou sémantique ? », *Langue francaise* 103, pp. 32-44.

ARONOFF, M. (1976), *Word Formation in Generative Grammar*, Cambridge, Mass. : MIT Press.

CORBIN, D. (2004), « French (Indo-European: Romance) », *in* G. BOOIJ, C. LEHMANN & J. MUGDAN (eds.), *An International Handbook on Inflection and Word Formation*, vol. 1, art. 121, New-York : Mouton - Walter de Gruyter.

DAL, G., & F. NAMER (2005), « L'exception infirme-t-elle la notion de règle ? Ou le lexique construit et la théorie de l'optimalité », *Faits de langue* 25, pp. 123-130.

FRADIN, B. (1997), « Esquisse d'une sémantique de la préfixation en *anti-* », *Recherches linguistiques de Vincennes* 26, pp. 87-112.

FRADIN, B. (2000), « Combining forms, blends and related phenomena », *in* U. DOLESCHAL & A.M. THORNTON (eds.), *Extragrammatical and Marginal Morphology*, München : Lincom Europa, pp. 11–59.

GAWELKO, M. (1977), *Evolution des suffixes adjectivaux en français,* Wrocław : Polska Akademia Nauk Komitet Neofilologiczny.

GREFENSTETTE, G. (1999), « The WWW as a Resource for Example-Based MT Tasks », *in Proceedings of the ASLIB 'Translating and the Computer' Conference*, London.

HABERT, B., A. NAZARENKO & A. SALEM (1997), *Les linguistiques de corpus*, Paris : Armand Colin / Masson.

HATHOUT, N., M. PLÉNAT & L. TANGUY (2003), « Enquête sur les dérivés en *-able* », *Cahiers de Grammaire* 28, pp. 49–90.

HATHOUT, N., & L. TANGUY (2002), « Webaffix : finding and validating morphological links on the WWW », *in Proceedings of the Third International Conference on Language Resources and Evaluation*, ELRA, Las Palmas de Gran Canaria, pp. 1799–1804.

IACOBINI, C. (2004), « Composizione con elementi neoclassici », *in* M. GROSSMANN & F. RAINER (eds.), *La formazione delle parole in italiano*, Tübingen : Niemeyer, pp. 69–96.

JACQUEMIN, C., & C. BUSH (2000), « Combining lexical and formatting clues for named entity acquisition from the Web », *in Joint SIGDAT Conference on Empirical Methods in NLP and Very Large Corpora*, Hong Kong, pp. 181–189.

LEEMAN, D. (1992), « Deux classes d'adjectifs en *-ble* », *Langue française* 96, pp. 44-64.

LEEMAN, D., & S. MELEUC (1990), « Verbes en tables et adjectifs en *-able* », *Langue française* 87, pp. 30-51.

LIEBER, R. (1983), « Argument Linking and Compounds in English », *Linguistic Inquiry* 14.2, pp. 251–85.

LOVIS, Chr., P.A. MICHEL, R. BAUD & J-R. SCHERRER (1995), « Word segmentation processing: a way to exponentially extend medical dictionaries », *in* R.A. GREENES, H.E. PETERSON & D.J. PROTTI (eds), *Proceedings of the 8th World Congress on Medical Informatics*, pp. 28-32.

LOVIS, Chr., R. BAUD, A-M. RASSINOUX, P.A. MICHEL & J-R. SCHERRER (1998), « Medical dictionaries for patient encoding systems: a methodology », *Artificial Intelligence in Medecine* 14, pp. 201-214.

MOLINIER, Chr. (1992), « Sur la productivité adverbiale des adjectifs », *Langue française* 96, pp. 65-73.

NAMER, F. (2002), « Valider les unités morphologiques par le Web », *in* B. FRADIN, G. DAL, N. HATHOUT, F. KERLEROUX, M. PLÉNAT & M. ROCHÉ (éds), *Silexicales 3: les unités morphologiques,* SILEX, Université de Lille III, Villeneuve d'Ascq, pp. 142–150.

NAMER, F. (2003a), « Productivité morphologique et complexité de la base : le système MoQuête », *Langue Française* 140, pp. 79–101.

NAMER, F. (2005), « Le modèle Lstat: ou comment se constituer une base de données morphologique à partir du Web », *Revue Québécoise de Linguistique* 32.1, pp. 85-110.

PICHON, E. (1940), « Attache d'un suffixe à un complexe », *Le Français moderne* 8, pp. 27-23.

PLÉNAT, M. (1988), « Morphologie de adjectifs en *-able* », *Cahiers de grammaire* 13, pp. 101–132.

PLÉNAT, M. (1997), « Analyse morpho-phonologique d'un corpus d'adjectifs en *-esque* », *Journal of French Language Studies* 7, pp. 163–179.

PLÉNAT, M. (2000), « Quelques thèmes de recherche actuels en morphophonologie française », *Cahiers de lexicologie* 77, pp. 27–62.

PLÉNAT, M., S. LIGNON, N. SERNA & L. TANGUY (2002), « La conjecture de Pichon », *Corpus et recherches linguistiques* 1, pp. 105-150.

RENOU, L. (1930), *Grammaire sanscrite. Tome 1: Phonétique, composition, dérivation*, Paris : Adrien Maisonneuve.

RESNIK, P. (1999), « Mining the Web for bilingual text », *in Proceedings of the 37th Meeting of ACL*, Maryland, USA, pp. 527–534.

SCHULZ, S., M. ROMACKER, F. PIUS, F. ZAISS, K. RÜDIGER & U. HAHN (1999), « Towards a multilingual morpheme thesaurus for medical free-text retrieval », *in Proceedings of Medical Informatics Europe (MIE)*, Ljubiana, Slovenia, pp. 891-894.

TANGUY, L., & N. HATHOUT. (2002). « Webaffix : un outil d'acquisition morphologique dérivationnelle à partir du Web », *in* J.-M. PIERREL (éd.), *Actes de la 9e Conférence Annuelle sur le Traitement Automatique des Langues Naturelles (TALN-2002)*, Nancy : ATALA, pp. 245–254.

VILLOING, F. (2003), « Les bases des opérations de construction morphologiques : des unités sémantiquement spécifiées. Illustration à la lumière de la composition $[VN]_{N/A}$ en français », *in* B. FRADIN, G. DAL, N. HATHOUT, F. KERLEROUX, M. PLÉNAT & M. ROCHÉ (éds), *Silexicales 3 : les unités morphologiques.*, SILEX, Université de Lille III, Villeneuve d'Ascq, pp. 213-219.

WARREN, B. (1990), « The importance of combining forms », *in* W.U. DRESSLER, H.C. LUSCHÜTZKY, O.E. PFEIFFER & J.R. RENNISON (eds.), *Contemporary Morphology*, Berlin, New York : Mouton - Walter de Gruyter, pp. 111-32.

WILLIAMS, E. (1981), « On the Notions 'Lexically Related' and 'Head of a Word' », *Linguistic Inquiry* 12.2, pp. 245-74.

YVON F. (1996), *Prononcer par analogie : motivation, formalisation et évaluation*, Thèse de doctorat de l'E.N.S.T, Paris.

ZWEIGENBAUM, P., R. BAUD, A. BURGUN, F. NAMER, E. JARROUSSE, N. GRABAR, P. RUCH, F. LE DUFF, J.-F. FORGET, M. DOUYERE, & S. DARMONI (2005), « UMLF: a unified

medical lexicon for French », *International Journal of Medical Informatics* 74.2-4, pp. 119-124.