Durand J. Habert B., Laks B. (éds.)
Congrès Mondial de Linguistique Française - CMLF'08
ISBN 978-2-7598-0358-3, Paris, 2008, Institut de Linguistique Française

Lexique(s)
DOI 10.1051/cmlf08332

Perspectives de la lexicographie informatisée

Robert Martin

Institut de France

1 Introduction

Les perspectives de la lexicographie informatisée semblent de deux ordres : d'une part la technique, qui a déjà, grâce à l'ordinateur, considérablement évolué, peut encore sensiblement progresser ; d'autre part, tout porte à croire qu'il faut rapprocher le plus possible la lexicographie à destination humaine et les "dictionnaires-machine" indispensables au TAL.

2 La technique lexicographique : des acquis aux perspectives

Dès à présent, l'informatique a profondément modifié les techniques lexicographiques. On présentera tout d'abord les avancées les plus notables, puis on évoquera les perspectives qui s'ouvrent, car divers aspects demeurent insuffisamment développés, notamment celui de la rédaction assistée par ordinateur.

2.1 Les avancées les plus marquantes

Parmi les avancées les plus marquantes, et qui sont désormais des acquis, une place importante revient

- à la documentation informatisée

- au balisage systématique des informations

- à l'essor d'une lexicographie évolutive

- au développement de l'hypertextualité lexicographique.

2.1.1 La documentation informatisée

L'informatisation de la documentation lexicographique s'impose désormais comme une technique indispensable. On n'imagine plus qu'un dictionnaire puisse s'élaborer sans que la documentation qui le fonde soit au moins en grande partie informatisée. En lexicographie française, le TLF a joué à cet égard un rôle déterminant. Dès 1964, grâce à l'action de Paul Imbs, le "Centre de recherche pour un Trésor de la langue française" (CRTLF) a bénéficié, comme on sait, d'un équipement tout à fait exceptionnel pour l'époque : un ordinateur "Bull-Gamma 60" a été installé à Nancy. Certes on était bien loin des techniques miniaturisées d'aujourd'hui : il a même fallu construire, pour loger le monstre, un bâtiment approprié, muni en sous-sol d'un système encombrant de climatisation. C'étaient les débuts ! Mais les services rendus ont très vite été remarquables, à la fois en quantité et en qualité. Voyez la bibliographie du TLF : la base des textes enregistrés, qui a ensuite formé FRANTEXT, a été constituée pour l'essentiel dès les années soixante. Dès le départ aussi, on a élaboré des instruments qui ont permis le regroupement des formes fléchies et le repérage des formes homographes. Le "dictionnaire des formes flexionnelles", qui remonte à 1965, est l'ancêtre de ce que l'on a appelé depuis "les analyseurs morphologiques". Par ailleurs, on s'est appliqué, à la fin des années soixante, à opérer dans cette immense masse de données des repérages sélectifs moyennant des procédures statistiques (notamment la "loi de Poisson" ; *Cahiers de lexicologie*, 1971).

Mais laissons-là le passé. Ce qui est certain, c'est que la lexicographie est indiscociable de l'aspect massif. Pas de dictionnaire en-deça d'un certain volume : il y faut une nomenclature de vaste étendue et des informations de grande richesse. Le traitement, même détaillé, d'une centaine de mots ne fera jamais un

Durand J. Habert B., Laks B. (éds.)
Congrès Mondial de Linguistique Française - CMLF'08
ISBN 978-2-7598-0358-3, Paris, 2008, Institut de Linguistique Française

Lexique(s)
DOI 10.1051/cmlf08332

"dictionnaire". Les vocables se comptent nécessairement par dizaines de milliers, voire par centaines. Bref, l'aspect cumulatif de la lexicographie s'impose comme une évidence. Cela revient à dire que la lexicographie est liée par nature aux techniques informatiques : seule l'informatique permet de traiter commodément les masses immenses de données que l'élaboration du dictionnaire requiert.

2.1.2 Le balisage des informations

Une évolution plus récente est celle des "dictionnaires informatisés". C'est là un point certes essentiel, mais que l'on n'évoquera que pour mémoire, la lexicographie informatisée ayant atteint désormais l'âge mûr et ses exigences étant bien connues. Là encore le TLF a joué un rôle déterminant. Son degré d'informatisation dépasse de beaucoup celui de son illustre prédécesseur, le Dictionnaire d'Oxford. Par ailleurs, le libre accès sur Internet a été une avancée considérable.

Il semble quasiment impossible qu'à l'avenir un grand dictionnaire, du moins scientifique (par opposition au dictionnaire commercial), puisse se passer d'une version informatisée. Les avantages de la consultation informatisée sont tels qu'il paraît difficile désormais de s'en priver. Déliée de la linéarité, elle confère à l'ouvrage une dimension comparable aux "3D" de l'architecture : c'est devenu de l'ordre de l'indispensable.

2.1.3 L'essor d'une lexicographie évolutive

Un dictionnaire, par nature, est un objet inachevé. A peine publié, les données qu'il comporte s'étant inévitablement accrues, une édition nouvelle devrait déjà se préparer. Voyez le DEAF : au fur et à mesure que ses (excellents) fascicules paraissent, les rédacteurs éprouvent le besoin d'en augmenter les données et d'en rectifier les insuffisances. Il arrive fréquemment ainsi que, pour tel ou tel vocable, il faille aller aussi au fascicule suivant, voire au suivant encore : avouons que ce n'est pas très commode, d'autant plus que l'on ne sait pas a priori s'il ne faut pas aller plus loin.

L'informatique résout élégamment de tels problèmes. L'article rédigé peut en effet être relié automatiquement aux ajouts postérieurs, à moins que l'on ne préfère tout simplement réécrire l'article (ou du moins y porter directement les ajouts et les corrections).

Informatisée, la lexicographie devient ainsi évolutive. Inutile d'attendre pour corriger ou pour augmenter une hypothétique réédition, elle-même périmée aussitôt qu'elle paraît : la matière peut en être indéfiniment remodelée.

Une question très importante cependant est de savoir si les retouches doivent être quasi-quotidiennes ou s'il vaut mieux procéder par éditions successives, par exemple annuelles. La seconde solution paraît de loin préférable, pour une raison toute simple : un objet en évolution constante ne peut servir de référence ; impossible de citer une telle source. Mieux vaut à tous égards des éditions datées, quitte à les enchaîner dans des délais rapprochés. C'est ce qui se fait pour le DMF. En 2003 a été mise sur la toile une première version (le DMF1, accumulation lemmatisée de 13 lexiques "préalables") ; une seconde version, considérablement enrichie, date de 2007 (le DMF2 ; cette version comporte notamment un "lexique complémentaire" qui recueille quelque 35000 vocables qui ne sont pas déjà dans le DMF1) ; une troisième version (le DMF3 synthétisant progressivement l'ensemble des articles et enrichissant les plus importants) devrait paraître, d'année en année, à partir de 2009. Au reste, l'ensemble des éditions demeure accessible, même celles qui sont dépassées. Ainsi conçue, quoique évolutive, la lexicographie présente aussi la stabilité que requiert l'édition scientifique.

2.1.4 L'hypertextualité lexicographique

Un autre avantage, capital, que suscite l'informatisation est l'accès hypertextuel. Le dictionnaire peut être mis en relation

- avec tous les autres dictionnaires informatisés (pour peu que leur lemmatisation soit compatible)

Durand J. Habert B., Laks B. (éds.)
Congrès Mondial de Linguistique Française - CMLF'08
ISBN 978-2-7598-0358-3, Paris, 2008, Institut de Linguistique Française

Lexique(s)
DOI 10.1051/cmlf08332

- avec les bases de données qui ont servi à le construire.

Ainsi, le TLF est mis en relation avec les dictionnaires que l'ATILF abrite et avec les dictionnaires de l'Académie ; par ailleurs un pont est jeté avec FRANTEXT (du moins avec les textes hors droits). Le DMF quant à lui est lié au TLF (et conséquemment aux liens du TLF) ; sa bibliographie est connectée à celle du DEAF ; bientôt un lien sera établé avec l'AND. Par ailleurs, un "double-clic" sur n'importe quelle forme, dans un exemple cité, déclenche sa lemmatisation et mène non seulement à l'article correspondant du DMF, mais à toutes les occurrences de la Base textuelle qui sous-tend le dictionnaire. Plus généralement, il est envisagé d'ouvrir par cette voie l'ensemble de la documentation du DMF (docDMF).

Ces deux exemples montrent la très grande importance de l'hypertextualité. Le CNRTL représente dans ce domaine une initiative particulièrement prometteuse.

2.2 Vers une rédaction assistée par ordinateur d'un côté et une consultation plus personnalisée de l'autre

Tout cela appartient (plus ou moins) à l'acquis. Les perspectives d'avenir sont plutôt ailleurs : du côté de la rédaction, dans une assistance accrue de l'ordinateur et, du côté de l'utilisateur, dans une consultation plus personnalisée.

2.2.1 Vers une rédaction assistée par ordinateur

L'expérience du DMF montre le gain que procure une rédaction assistée par ordinateur, même si d'importants perfectionnements restent envisageables.

Grâce à Gilles Souvay (ATILF) qui a conçu et réalisé les programmes nécessaires, la rédaction du DMF s'effectue au moyen de deux instruments déjà fort efficaces : d'une part un "masque de saisie" ; d'autre part un "correcteur lexicographique".

Tout dictionnaire se fonde sur une grammaire rédactionnelle : les informations se rangent par catégories (vedette, code grammmatical, définition, exemple...), et ces informations se suivent dans un ordre déterminé. Le "masque de saisie" comporte l'ensemble des balises et les valeurs qu'elles peuvent prendre (p. ex. la balise DOM[aine] peut prendre la valeur ASTR., MAR., MéD. ... ; la balise INDIC[ateur] les valeurs p. ext., p. méton., p. anal., au fig. ... ; la balise CODE GR[ammatical] les valeurs subst. masc., subst. fém., adj., verbe ...) ; par ailleurs à chaque endroit du texte lexicographique où le rédacteur est parvenu, le "masque" indique les possibles autorisés (p. ex. après le lemme vient obligatoirement le code grammatical, après le numéro de paragraphe se place nécessairement ou bien l'indicateur des "conditions d'emploi", ou bien un syntagme (construction ou locution), ou bien la définition ; on peut ne remplir qu'un seul de ces champs ; si plusieurs le sont, c'est forcément dans l'ordre que l'on vient d'indiquer. Bref, à chaque étape, le rédacteur est guidé dans ses choix ; la structure du dictionnaire est garantie par le système. C'est dire en même temps que le balisage est entièrement cohérent : plus de procédure coûteuse de "rétroconversion" pour informatiser le dictionnaire (comme pour le TLF) : le dictionnaire est structuré et balisé dès la rédaction.

Au "masque de saisie" s'ajoute un "correcteur lexicographique". Sa finalité est de vérifier divers aspects. Il vise notamment :

- le repérage d'erreurs matérielles (p. ex. le correcteur indique si le rappel des entrées de certains dictionnaires, dont la nomenclature (et la tomaison) a été enregistrée, est correcte, ainsi pour le T-L ou pour le FEW ; ou bien si la suite de la numérotation est cohérente ; ou encore si les exemples figurent bien dans l'ordre chronologique sous chacune des rubriques...) ;

- le contrôle des valeurs attribuées aux balises (p. ex, sous DOM., le correcteur vérifie que la valeur appartient bien aux domaines enregistrés) ;

- le contrôle des références bibliographiques, etc.

Durand J. Habert B., Laks B. (éds.)
Congrès Mondial de Linguistique Française - CMLF'08
ISBN 978-2-7598-0358-3, Paris, 2008, Institut de Linguistique Française

Lexique(s)
DOI 10.1051/cmlf08332

Toutes ces techniques apportent dès à présent une aide considérable. Mais c'est assurément dans ce domaine que l'évolution reste le plus largement ouverte. Le progrès peut aller dans diverses directions, p. ex. :

- il serait intéressant de fusionner "masque de saisie" et "correcteur" ; ainsi le système n'accepterait, sous la balise DOM., que les valeurs qu'il connaît (p. ex. en les affichant) ; en fait le problème est plus délicat qu'il n'y paraît ; il faut en effet qu'au fil de la rédaction, l'article puisse être constamment remodelé ; une grammaire trop contraignante risquerait de perturber l'opération en exigeant à tout moment une parfaite cohérence. L'équilibre entre l'évolution de l'article, parfois complètement à remanier, et l'indispensable complétude de la structure finale est assurément difficile à réaliser.

Un autre sujet de réflexion porte sur la meilleure façon de corriger et d'enrichir le dictionnaire. Un article devrait pouvoir être extrait facilement à tout moment pour être mis en réserve du prochain "montage" (l'édition suivante) ; cette réserve elle-même devrait rester accessible au rédacteur pour d'éventuels nouveaux ajouts. Il y a là une gestion délicate qu'il y a tout à gagner à étudier en détail.

Bref, les progrès à réaliser dans ce domaine restent nombreux, en dépit d'avancées tout à fait appréciables.

2.2.2 Vers une consultation plus personnalisée

Des perspectives prometteuses sont envisageables aussi du côté de l'utilisateur. Le point essentiel est de personnaliser la consultation. On peut envisager au moins deux directions :

- Actuellement les dictionnaires informatisés ne font pas la place qu'il faudrait à l' "annotation" ; les données du TLF sont certes transférables (hors balisage, bien évidemment) ; mais il serait très utile que le consultant puisse se constituer une base personnelle activable à chaque accès au TLF ; il disposerait, si l'on préfère, d'un TLF remodelé à sa façon ; naturellement ces données ne seraient pas stockées sur le site du dictionnaire, TLF ou autre, mais sur celui du consultant, un programme spécifique permettant de reconstruire l'état auquel, article modifié par article modifié, la matière est parvenue. Vaste programme ! Mais ce n'est tout de même pas de l'ordre de l'inimaginable.

- Un autre développement consisterait à promouvoir une lexicographie "modulable" ; comme on sait, il existe toujours, sous une même entrée, un grand nombre de possibles pour en traiter (même à l'intérieur d'une "grammaire" donnée) ; l'idée serait de permettre, lors de la consultation, de modifier les critères adoptés et de reconstruire l'article selon un plan différent. On renvoie sur ce point au *Bull. Soc. Ling. Paris* 102, 2007, 17-33.

3 La jonction de la lexicographie à destination humaine et les "dictionnaires-machine"

Si donc les techniques sont susceptibles de progresser, il est une autre perspective qui est également, semble-t-il, de grande importance. Le dessein serait de rapprocher le plus possible la lexicographie à destination humaine et les "dictionnaires-machine" nécessaires au TAL. Les deux types y gagneraient. La lexicographie "humaine" en rigueur ; la lexicographie "mécanique" en richesse. Aucun dictionnaire "humain" n'a la cohérence nécessaire pour se prêter directement à l'exploitation du TAL : il doit être considérablement réaménagé. Aucun "dictionnaire-machine" ne comporte un volume d'information comparable à celui du dictionnaire "humain" : ses données sont certes correctement formalisées, mais en général elles restent désespérément pauvres. On peut donc raisonnablement penser qu'une voie d'avenir est de rapprocher au mieux les deux types. En linguistique française, on est très loin du compte. Il existe un dictionnaire du coréen contemporain qui réalise peu ou prou cette exigence : il a été présenté par ses concepteurs de l'Université Nationale de Séoul, Seong Heon Lee et Chai-Song Hong, au Colloque qui s'est tenu, à l'initiative de l'ATILF, en janvier dernier (Pré-Actes, 159-165). L'allure de ce dictionnaire paraîtra sans doute un peu raide, si ce n'est rébarbative ; mais il est certain qu'il est utilisable indifféremment par l'homme et par la machine.

Durand J. Habert B., Laks B. (éds.)
Congrès Mondial de Linguistique Française - CMLF'08
ISBN 978-2-7598-0358-3, Paris, 2008, Institut de Linguistique Française

Lexique(s)
DOI 10.1051/cmlf08332

Pour remplir au mieux une telle finalité, tout en conservant à l'objet un aspect qui ne soit pas décourageant, il convient pour le moins, sans doute en conservant des zones d'écriture libre placées en dehors de l'exploitation mécanique, de veiller à deux aspects : une écriture aussi unifiée que possible ; des formulations parfaitement explicites.

3.1 Une écriture unifiée

La première exigence est celle d'une écriture unifiée. Il arrive constamment, en lexicographie, qu'un même type d'information se présente sous des formes différentes.

Ainsi pour les informations grammaticales. L'emploi pronominal d'un verbe peut être signalé par l'indicateur grammatical "Empl. pronom." ; il peut aussi être exhibé, surtout si la construction demande à être spécifiée (*s'en remettre à qqn*) ; le mieux sans doute serait de donner dans tous les cas les deux (Empl. pronom. *S'en remettre à qqn*). Une telle présentation unifiée n'est pas une gêne, au contraire, pour le consultant humain ; et elle a l'avantage de rester ouverte à l'exploitation mécanique. Moyennant une écriture unifiée, on donne aux objets une allure qui les rend compatibles avec la représentation formalisée qu'exige le traitement automatique.

Autre cas, celui-ci plus sémantique. La nature sémantique de l'entourage, sous un sens donné, peut être spécifiée de trois manières

- à l'intérieur de la définition : *remettre* "appliquer à nouveau ou en plus (un produit de beauté, d'hygiène ou de soin)"

- dans la construction exhibée : *remettre* (un produit de beauté, d'hygiène ou de soin) "L'appliquer à nouveau ou en plus".

- ou encore dans les "conditions d'emploi" : [Le compl. d'obj. désigne un produit de beauté, d'hygiène ou de soin] "L'appliquer à nouveau ou en plus"

Il serait judicieux d'opter là aussi pour une présentation unifiée. Peu importe la redondance, p. ex. *Remettre* (un produit de beauté, d'hygiène ou de soin) "Appliquer à nouveau ou en plus (un produit de beauté, d'hygiène ou de soin)". L'essentiel est que la présentation soit partout la même.

On multiplierait à l'infini les exemples. Au reste, l'unification touche le dictionnaire comme objet particulier. Mais on peut imaginer, plus généralement, l'adoption de règles communes ; ce n'est pas une mince affaire ; mais le bénéfice en serait considérable.

3.2 Des formulations explicites

Dans cet effort pour concilier, autant que faire se peut, la lexicographie "humaine" et la lexicographie "mécanique", une autre préoccupation est celle de l'explicitation. Rien de pire en matière de TAL qu'une écriture allusive ou ambigüe.

Voyez p. ex. la présentation des constructions :

Publication d'une ordonnance, d'un acte de mariage se résout ainsi : *publication d'une ordonnance / publication d'un acte de mariage*

Proclamation du chef, de l'acte d'accusation se résout par : *proclamation du chef d'accusation / proclamation de l'acte d'accusation*.

Pour éviter **publication d'une ordonnance de mariage* sur le modèle du deuxième cas, mieux vaut adopter partout l'écriture explicite qui suit le double point. C'est un peu plus lourd à la lecture ; mais c'est pertinent dans tous les usages.

L'exigence d'explicitation a par ailleurs une portée beaucoup plus générale : la lexicographie de l'avenir, plus que celle que nous connaissons, s'appliquera assurément, non seulement à spécifier au mieux les

Durand J. Habert B., Laks B. (éds.)
Congrès Mondial de Linguistique Française - CMLF'08
ISBN 978-2-7598-0358-3, Paris, 2008, Institut de Linguistique Française

Lexique(s)
DOI 10.1051/cmlf08332

acceptions, mais encore à préciser les conditions qui doivent être satisfaites contextuellement pour leur émergence. La notion de "condition d'emploi" initiée par le TLF va dans ce sens. Mais dans cette perspective, il reste beaucoup faire. Les dictionnaires de l'avenir, tout donne à le penser, auront vraisemblablement une allure assez différente des nôtres. Ils s'appliqueront certes à dire le sens des mots mais ils expliciteront aussi les conditions auxquelles les sens se réalisent.

On n'évoque pas l'avenir en dehors de la conjecture : rien d'assuré dans tout cela. Que sera la lexicographie de demain ? Personne ne peut le prédire. Une chose pourtant semble à peu près certaine : les rapports de la lexicographie (scientifique) et de l'informatique ont toute chance de devenir de plus en plus étroits.