3

NOUVELLES TECHNOLOGIES, NOUVEAUX MODÈLES LINGUISTIQUES ET NÉOLOGIE

Emmanuel CARTIER et Jean-François SABLAYROLLES
Université Paris 13, Laboratoire LDI (UMR 7189)

INTRODUCTION

Le lexique et les dictionnaires occupent une place centrale en TAL. Parmi les modèles descriptifs récents en ce domaine, celui des classes d'objets développé par Gaston Gross au LDI – avec la confection de dictionnaires de prédicats et d'arguments explicitant la combinatoire et l'actualisation possibles de tous les éléments de chaque classe – a des incidences sur la conception de la néologie et l'extraction des néologismes. Par ailleurs, les bases de données informatisées permettent non seulement d'engranger des informations, comme les fiches cartonnées d'antan, mais elles sont aussi de véritables outils de recherche qui, par des requêtes simples ou croisées, permettent une exploitation directe et efficace des données pour une meilleure connaissance du phénomène, et assurent aussi le suivi de la diffusion des néologismes.

1. INCIDENCES SUR LA CONCEPTION DE LA NÉOLOGIE

Contrairement à la doxa, ce n'est pas le concept de néologie qui est flou, c'est l'absence de descriptions systématiques de la langue qui pose problème dans le repérage des néologismes : nous ne disposons pas de pierres de touche fiables pour décider de l'existence d'une innovation lexicale, qu'il s'agisse de l'apparition d'une nouvelle lexie [1], inexistante auparavant ou d'un emploi innovant par

1. Sous ce terme, créé par B. Pottier, et repris par d'autres linguistes dont J. Tournier, comme dénomination de l'unité lexicale, nous entendons un signe linguistique ayant une unité fonctionnelle et mémorisée (ou mémorisable) en compétence.

rapport à la combinatoire attestée d'une unité déjà existante dans la langue. Le recours au(x) dictionnaire(s) traditionnel(s) comme corpus d'exclusion ne peut être mécanique du fait de leurs différences, des lacunes de leurs nomenclatures, de l'insuffisance de leurs descriptions et de l'existence, généralement oubliée, des néologismes homonymiques [2]. Bien sûr ces dictionnaires sont des outils qui apportent nombre d'informations intéressantes, mais leur manque de systématicité les rend difficilement utilisables. C'est là où réside la supériorité des dictionnaires directement informatisés et comportant des informations sur les divers emplois d'un signifiant [3]. Une lexie trouvée dans un texte et absente comme entrée ou avec un emploi non décrit relève de la néologie, sauf en cas d'une éventuelle lacune du dictionnaire, qu'il faut alors modifier en conséquence. Outre les emprunts et la néologie formelle (morpho-sémantique ou purement morphologique de J. Tournier [4]), il y a la néologie syntactico-sémantique [5] et la néologie syntaxique (des changements de construction sans modification sensible du sens [6]).

2. Sur les rapports entre les néologismes et les dictionnaires, voir entre autres Sablayrolles (2002, 2008).

3. On évite volontairement d'employer ici le terme *lexie*, dans la mesure où cela entraînerait une longue discussion sur le traitement polysémique ou homonymique des lexies.

4. Outre les procédés traditionnels et bien connus de la néologie morpho-sémantique que sont la dérivation et la composition sous leurs diverses formes, il y a, moins connues, les innovations flexionnelles sans changement de sens (*je closis*, *la représaille*) ou avec des inflexions de sens plus notables (*les banlieues* ou *les quartiers* qui ne s'appliquent qu'à des lieux défavorisés ou à problèmes). La néologie purement morphologique correspond aux diverses troncations et siglaisons.

5. Entrent dans cette grande catégorie tous les changements d'emploi qui ne s'accompagnent pas d'une modification affixale : les conversions et la néologie sémantique. Dans les deux cas, il y a infraction par rapport à la combinatoire décrite systématiquement dans des dictionnaires informatiques, ainsi de l'emploi maintenant répandu de l'adjectif *grave* comme adverbe (*ça m'ennuie grave*) ou de tous les emplois figurés comme *formater* pour un humain (*Jean a été formaté pour ce poste*).

6. Ce concept et cette dénomination sont anciens et se trouvent exposés, entre autres, dans le *Dictionnaire Universel* de Pierre Larousse et au détour de certaines pages de l'*Histoire de la Langue française* de Ferdinand Brunot. On peut en distinguer au moins deux grandes catégories. La première consiste essentiellement dans des changements dans l'emploi de préposition (suppression : *il craint*, en français non standard, ajout : *pallier à*, remplacement par une autre : *semblable avec*). La seconde se manifeste dans l'apparition d'un prédicat sous une forme catégorielle inattendue : si la forme verbale *prendre* et la forme nominale *prise* sont possibles quand il s'agit des classes d'objets <médicament> ou <lieu institutionnel> (*prendre un médicament* / *la prise de ce médicament* ; *prendre Troie*, *la Bastille* /*la prise de Troie*, *de la Bastille*), seule la forme verbale est normalement attestée pour les <moyens de transports> : *prendre le bus* / **la prise du bus*. C'est donc avec une innovation (par transgression volontaire destinée à frapper le public cible) que des publicitaires ont créé le slogan au bénéfice des TER de la SNCF *La prise de train bénéficie à la santé de votre voiture*.

La néologie par détournement pose des problèmes spécifiques qui ne seront pas abordés ici.

2. INCIDENCES PRATIQUES : L'EXTRACTION AUTOMATIQUE DES NÉOLOGISMES

L'extraction (semi)automatique des néologismes se fait à l'aide de corpus d'exclusion. Le problème principal repose donc sur la qualité et la fiabilité de ce corpus d'exclusion. S'il se contente de reprendre la nomenclature des dictionnaires papier, on a les mêmes problèmes que ceux déjà évoqués ci-dessus et ailleurs. Il faut donc se confectionner un dictionnaire informatique plus fiable ou en avoir un à sa disposition. C'est le cas au LDI avec le dictionnaire *Morfetik* élaboré depuis des années par Michel Mathieu-Colas. Il comporte de l'ordre d'un million de formes et le dictionnaire des mots composés, encore indépendant, va y être incessamment intégré. À l'aide de l'outil *Telanaute* mis au point par Fabrice Issac, nous importons des textes de la presse quotidienne en ligne dont les « mots » sont confrontés au dictionnaire *Morfetik*. Tout ce qui n'est pas reconnu est présenté comme candidat au néologisme. Trois situations se présentent alors. Soit il s'agit de vrais néologismes, et ils sont intégrés dans la base *Neologia* (voir *infra*), soit ce sont des lacunes de *Morfetik* et ces lexies sont intégrées dans le dictionnaire, soit, et c'est quantitativement encore le plus fréquent, il s'agit d'erreurs diverses qui sont éliminées d'un coup de souris, mais qui sont intégrées dans le corpus d'exclusion afin de ne pas les voir réapparaître. Petit à petit, le nombre de ces scories tendra à diminuer sans pouvoir être totalement éliminées néanmoins.

Si le système d'extraction des néologismes formels est au point et fonctionne assez bien, l'extraction des autres néologismes, par changement de combinatoire (syntactico-sémantique ou purement syntaxique, voir *supra*), en est encore à ses balbutiements. Les processus de comparaison de la combinatoire effectivement employée dans les textes avec celle, conventionnelle, telle qu'elle devrait être consignée dans les dictionnaires de langue et telle que s'efforcent de la décrire les dictionnaires informatiques élaborés au LDI, sont envisageables dans l'avenir. Cependant, les extractions sont beaucoup plus complexes à mettre en œuvre, ne serait-ce que parce que les dictionnaires informatiques sont des dictionnaires de phrases minimales [7] et que les textes se présentent avec des phrases de structures plus complexes comprenant des prédicats de second degré, des expansions, des insertions, des problèmes d'ordre des

7. Une phrase minimale est un prédicat saturé par ses arguments.

constituants… Une phase de transition consiste donc à sérier les problèmes et à opérer sur des phrases réduites aux constituants obligatoires.

3. UNE BASE DE TRAITEMENT ET DE SUIVI DES NÉOLOGISMES, *NEOLOGIA*

L'informatique permet de replacer la néologie dans l'ensemble des phénomènes linguistiques et de modéliser un système capable de suivre la « vie » des néologismes, comme le montre la figure 1.

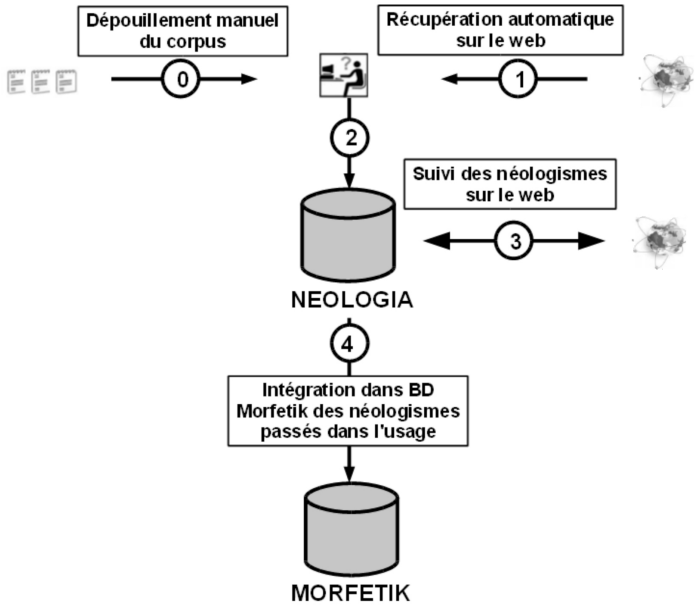

Figure 1.- Architecture générale d'un système de suivi des néologismes

Une base de données de néologismes, au sens informatique mais également dans un sens théorique, doit être replacée dans une architecture qui rende compte de toutes les étapes de la vie de ces éléments lexicaux :

- récupération automatique ou manuelle des néologismes (0), (1) ;
- saisie et structuration de l'information attachée à chaque néologisme (2) ;
- observation de la « vie » des néologismes (3) ;
- transfert des néologismes passés dans l'usage dans une base de données lexicographiques générale (4).

Nous détaillerons ci-après les étapes (2) et (3), (4).

Engranger les « prises » de la veille néologique mais aussi analyser les néologismes relevés, tels sont les deux objectifs majeurs de la base de donnée *Neologia*, mise au point par Emmanuel Cartier sur la base des champs informatifs proposés par Jean-François Sablayrolles. Cette base a d'emblée été conçue comme un véritable outil de gestion et d'étude des néologismes du français contemporain. Elle comporte les propriétés ergonomiques suivantes :

- Accès protégé avec trois niveaux :
  – Invité (consultation des fiches validées, sur demande),
  – Auteur (créateur de fiches),
  – Administrateur (validation de toutes fiches).
- Deux écrans principaux simples et un *gestionnaire* :
  – Création / édition des fiches,
  – Interface de recherche,
  – (Administrateur : paramétrage des champs).

Les entrées sont décrites au moyen d'une trentaine de champs répartis en cinq groupes : entrée et définition, morphosyntaxe, syntactico-sémantique, néologie, relations sémantiques et contextes présentés sous forme d'onglets [8]. L'interface se présente ainsi :

Figure 2.- Prise d'écran de l'interface création / modification de fiche

8. Pour plus de détail sur les différents champs d'information, v. Cartier & Sablayrolles (2009).

Les fonctionnalités d'édition de recherche et de consultation implantées dans l'outil permettent des requêtes simples (par matrices, par catégories grammaticales…) et des requêtes multicritères (associant deux, trois critères ou plus) de divers types.

Le pouvoir heuristique de cette base sur le travail n'a pas été négligeable. Les discussions lors de la création de la base ont conduit à approfondir la réflexion théorique dans le domaine en obligeant à expliciter jusque dans le détail les objectifs, et son utilisation a permis de découvrir des problèmes auxquels on n'avait pas songé initialement ; ce qui a également contribué à des approfondissements théoriques ou à des précisions rassemblés dans un recueil à usage interne de l'équipe, mais qui ont fait et feront l'objet de communications et d'articles.

4. SUIVI DE LA CIRCULATION DES NÉOLOGISMES JUSQU'À LA PERTE DE LEUR STATUT DE NÉOLOGISME

L'informatique fournit un outil pour gérer le cycle de vie de ces mots néologiques : émergence, développement ou disparition, intégration ou non dans les dictionnaires généraux. Il existe en effet des outils pour suivre le vocabulaire des corpus numériques : repérage des néologismes dans les corpus, enregistrement des premières occurrences, suivi de l'utilisation des mots, au moyen de différents outils statistiques sur le web. De la sorte, nous appliquons un modèle permettant de gérer le dynamisme de la nomenclature linguistique.

C'est ainsi que nous avons mis au point un outil permettant de gérer le cycle de vie des néologismes, en suivant son évolution au travers des flux textuels accessibles sur Internet à travers les grands moteurs de recherche. Ce suivi permet ensuite de profiler chaque néologisme et de définir un seuil au-dessus duquel un néologisme doit intégrer le dictionnaire général ou spécifique d'une langue.

CONCLUSION

Cet article a voulu démontrer à quel point l'informatique apporte des outils permettant d'amender et d'approfondir les conceptions linguistiques, en créant une obligation de mise en pratique salutaire. De ce point de vue, le travail que nous effectuons depuis maintenant presque cinq ans permet d'affirmer que la combinaison de l'informatique et de la linguistique est nécessaire et fructueuse.

Notre projet d'étude des néologismes est maintenant opérationnel, l'architecture en est clairement établie, mais, évidemment,

chaque niveau apporte son lot de difficultés : qualité des repérages de néologismes, actuellement limités aux termes restant après élimination des mots « connus » et ceux appartenant à un corpus d'exclusion ; affinement des informations attachées à chaque entrée ; affinement des procédures de suivi de néologismes.

RÉFÉRENCES BIBLIOGRAPHIQUES

CARTIER Emmanuel et SABLAYROLLES Jean-François, 2008, « Néologismes, dictionnaires et informatique », *Cahiers de Lexicologie*, 93, p. 175-192.

SABLAYROLLES Jean-François, 2002, « Fondements théoriques des difficultés pratiques du traitement des néologismes », *Revue française de linguistique appliquée*, 7(1), « Lexique : recherches actuelles », p. 97-111.

SABLAYROLLES Jean-François, 2008, « Néologie et dictionnaire(s) comme corpus d'exclusion », dans J.-F. Sablayrolles (éd.), *Néologie et terminologie dans les dictionnaires*, Paris, Champion, collection « Lexica », p. 19-36.