TTR
Traduction, terminologie, re?daction

Stratégie pour la détection semi-automatique des néologismes de presse

Maria Teresa Cabré and Lluís de Yzaguirre

Volume 8, Number 2, 2e semestre 1995

Technolectes et dictionnaires

URI: https://id.erudit.org/iderudit/037219ar
DOI: https://doi.org/10.7202/037219ar

See table of contents

Publisher(s)

Association canadienne de traductologie

ISSN

0835-8443 (print)
1708-2188 (digital)

Explore this journal

Cite this article

Cabré, M. T. & de Yzaguirre, L. (1995). Stratégie pour la détection semi-automatique des néologismes de presse. *TTR*, *8*(2), 89–100. https://doi.org/10.7202/037219ar

Article abstract

A Strategy for the Semi-Automatic Detection of Press Neologisms – The Observatori de Neologia de Barcelona (OBNEB) was created in order to automatize the detection of formal neologisms found in the press. In this article the basic concepts related to neology are introduced together with the process carried out at the OBNEB. The features of the Corpus of the University of Barcelona are presented in detail. Composed of seven millions of occurrences taken from press texts, this corpus was used to test a programme for the semi-automatic extraction of neologisms. Finally, the authors show how the programme developed at the Institut de Lingüística Aplicada manages to fulfill this task. Further information about the authors and their work can be found in "http.V/www.iula.upf.cs".

Stratégie pour la détection semi-automatique des néologismes de presse

Maria Teresa Cabré
Lluís de Yzaguirre

1. Le projet de base: Obnub (Observatori de Neologia)

1.1 Présentation du cadre du projet

L'observatoire de néologie de l'Université de Barcelone (Obnub) a été créé en 1988 dans le but de détecter des néologismes dans la presse. Ce travail sert à la recherche, à l'amélioration des dictionnaires et à l'aménagement linguistique. Il a été financé par l'Institut d'Estudis Catalans (qui détient l'autorité linguistique dans le domaine de la langue catalane) et par la Fundació Enciclopèdia Catalana (maison d'édition de dictionnaires en catalan). En janvier 1995, le siège de l'Obnub a été transféré à l'*Institut Universitari de Lingüística Aplicada de la Universitat «Pompeu Fabra»*; il a alors pris le nom d'«Obneb» (Observatori de Neologia de Barcelona), mais les liens qu'il entretient avec l'Université de Barcelone demeurent toujours les mêmes.

1.2 Le concept de néologisme lexical

La néologie lexicale comprend les aspects sémantiques, fonctionnels et formels; cela veut dire qu'il ne s'agit pas exclusivement de trouver des mots nouveaux, mais aussi des déplacements de signifiés qui sont enregistrés dans les dictionnaires (restriction, changement ou amplification du signifié) et des modifications dans le jeu fonctionnel d'un mot (la rection verbale, la concordance...).

1.3 Identification des néologismes

Pour identifier un mot en tant que néologisme ou repérer le type de changement qu'il a subi, on consulte un corpus d'exclusion, constitué par les principaux dictionnaires catalans et castillans (voir l'Annexe 1).

1.4 La néologie de large diffusion vs la néonymie

On recherche les néologismes dans la presse générale, parce que le principal objectif consiste à identifier les néologismes qui ont obtenu droit de cité dans des textes standards et/ou qui sont employés par un grand nombre de sujets parlants. On exclut préalablement toute nouveauté terminologique qui appartient à un domaine spécifique de la technologie, de la science ou de l'industrie.

2. Organisation du déroulement du travail à l'Obneb

2.1 Détection

Jusqu'à maintenant, la détection des phénomènes néologiques se faisait «manuellement», c'est-à-dire en se fiant à la compétence lexicale des collaborateurs et collaboratrices qui lisaient les journaux en y cherchant tout ce qui était intéressant pour la recherche; quand ils croyaient qu'un mot présentait une certaine nouveauté, ils le confrontaient avec le corpus d'exclusion lexicographique et, si un caractère nouveau se confirmait, ils remplissaient une fiche dont la structure et un exemple sont présentés à l'Annexe 2.

2.2 Collecte

Les fiches des néologismes sont vérifiées par des personnes différentes de celles qui les ont rédigées et, si nécessaire, discutées lors de sessions collectives. Si la vérification est positive, la fiche est introduite dans une base de données qui constitue le fichier de dépouillement. Celui-ci est structuré par langue et par période. De 1989 à 1994, 35 000 données ont été recueillies pour le catalan et 22 500 pour le castillan.

2.3 Du fichier de dépouillement au fichier d'analyse

Des chercheurs appartenant à l'*Observatori* analysent chaque fiche du fichier de dépouillement et y ajoutent des informations et des avis sur l'intérêt qu'il y a à conserver le néologisme potentiel. Dans cette procédure de vérification, les étapes sont:

a. L'identification.
b. L'analyse linguistique.
c. L'analyse sociolinguistique.
d. Les relations et les équivalences interlinguistiques.
e. Les résultats de la synthèse.
f. La diffusion.

3. Le projet électronique: du corpus textuel au logiciel d'analyse

3.1 Justification

Le projet Obneb, tel qu'il a été entrepris en 1988, continue à fonctionner, mais il exige d'importants coûts en ressources humaines. Nous avons donc décidé qu'il faudrait trouver des modes de recherche et de fonctionnement moins onéreux du point de vue du personnel et, si possible, plus systématiques et plus fiables. Nous avons considéré qu'il serait intéressant d'explorer les possibilités d'exploiter des corpora de presse en recourrant à la localisation automatique, d'abord pour le repérage des néologismes formels, puis pour l'identification des néologismes fonctionnels.

3.2 Le corpus textuel CECA

Au même moment, le Département de Philologie Catalane de l'Université de Barcelone décidait de la création d'un corpus représentatif du catalan non littéraire écrit et parlé d'aujourd'hui[1]. Dans ce corpus, appelé CECA, pour «Corpus Escrit del Català Actual», il y a une section réservée à la presse.

---

1. Le projet *Variació en el llenguatge : corpus oral i escrit de català contemporani* a été financé par la CIRIT (CS93-1017) et la DGICYT (PB 90-0505).

Type de corpus: Il s'agit de 119 journaux complets, qui vont du 17 février au 25 juin 1993. La publicité ajoutée en phase de maquettage n'y figure pas. Quelques jours manquent, et cela en raison de difficultés incontrôlables dans le système informatique du journal.

Acquisition: Les matériaux ont été transférés sur un PC qui communique avec l'ordinateur central au moyen d'un câble de type «null modem» et grâce à l'aide d'un logiciel de communication programmable. Cette technique a rendu possible l'automatisation de la réception de plusieurs numéros quotidiens (intégrés dans un fichier pour chaque page), et cela en une seule opération.

Stockage: Les quelque 7 000 fichiers, en format de photocomposition, ont été stockés dans un lecteur «Kodak WritableCD» et mis dans un CD-ROM pour mieux les conserver et pour en faciliter le transport et la manipulation.

Épuration: Puis, on a analysé les codes de photocomposition et développé un logiciel qui transforme l'ensemble des fichiers, un par un, en des fichiers ASCII. On a cependant conservé une partie des codes de photocomposition recodifiés, notamment ceux qui sont utiles à nos objectifs et qui facilitent le repérage de la date, de la page, de la section, de l'auteur, de la structure du texte (avant-titre, titre, sous-titre, paragraphe...), etc.

3.3 Analyse lexicale

a) *La désambiguïsation.* Les éléments ambigus sont des occurrences pour lesquelles l'ordinateur a plus d'un lemme possible, tandis que les néologismes formels sont des mots pour lesquels l'ordinateur n'a pas d'information. C'est une phase qui n'est donc pas strictement nécessaire pour la localisation des néologismes formels. Mais pour le néologisme fonctionnel, il faut une analyse et une compréhension de la phrase. Nous avons alors décidé, bien que le repérage des néologismes fonctionnels ne soit pas l'objectif principal du travail actuel, d'intégrer cette étape dans le processus de préparation du texte pour la localisation du néologisme formel.

b) *La localisation.* Pour des raisons d'efficacité, on localise d'abord les mots vides.

c) *La lemmatisation.* Pour cette étape, nous disposons d'un dictionnaire électronique de 68 514 lemmes qui se développe en 596 718 formes (avec les catégories morphologiques), dont les trois quarts sont des formes verbales; toutes les occurrences de mots du corpus qui se trouvent sur cette liste sont considérées comme des mots connus.

d) *La production des index.* Les mots du corpus qui ne se trouvent pas sur la liste des lemmes ou des formes dérivées sont cherchés dans des listes auxiliaires de formes onomastiques (patronymiques, toponymiques, etc.) ou de formes normalisées (formes abrégées, symboles, etc.).

e) *Les erreurs.* Si une occurrence n'a pas été résolue dans les phases précédentes, il est très possible qu'il s'agisse d'une erreur. L'analyse d'erreurs commence par la recherche du mot inconnu dans un fichier d'erreurs détectées au cours de contrôles antérieurs, parce que les erreurs se répètent souvent. Si le mot ne se trouve pas dans le fichier d'erreurs, la difficulté est résolue manuellement par le chercheur et le résultat est consigné au fichier. Afin d'éviter ces problèmes, on est en train d'élaborer une stratégie de repérage des erreurs basée sur une méthode «soundex» adaptée au catalan. Cette méthode consiste à rechercher des mots qui se ressemblent (permutations, suppressions, additions, altérations de lettres...) puis à les ordonner non seulement en fonction de leur proximité phonétique, mais aussi en s'appuyant sur leur probabilité d'apparition dans le corpus.

f) *Les cas non résolus.* Pour les mots non encore résolus, on est en train d'améliorer un analyseur morphologique, qui performe actuellement avec 80 % de succès. Si l'analyseur valide une recherche sur un mot, il s'agit probablement d'une unité parfaitement recevable, mais qui n'est pas consignée dans les dictionnaires parce qu'elle est automatiquement prédictible, comme c'est le cas de la plupart des superlatifs. Lorsque le mot est validé par un chercheur de l'équipe, il est ajouté au dictionnaire électronique.

g) *Les néologismes.* Les mots qui ont résisté à toutes les analyses décrites dans les phases précédentes sont considérés comme des

néologismes. Ils seront passés au crible par le personnel de l'«Observatori».

Il faut considérer que toute cette démarche d'analyse suit une progression croissante parce que chaque décision prise sur un mot enrichit le corpus de données nouvelles et qu'elle sert à établir de nouveaux critères, de telle sorte qu'on peut effectivement traiter un plus grand nombre de journaux avec le même nombre de personnes dans l'équipe, tout en augmentant la qualité des résultats obtenus.

4. Perspectives d'avenir

Nos travaux ont bien progressé jusqu'à présent, mais il reste toujours beaucoup à faire. Nous continuons nos recherches afin d'améliorer les méthodes de détection semi-automatique. Par ailleurs, nous avons également commencé à préparer d'autres travaux:

- Édition d'index et de concordances.

- Production de dictionnaires de néologismes.

- Production de textes dans plusieurs formats: TACT, TEI (avec lemmes et marquage morphologique), pre-MACHINAL...

- Généralisation d'outils d'analyse et de traitement.

- Automatisation du procédé de détection et de production des fichiers.

- Développement de l'analyse syntaxique pour la détection de la néologie fonctionnelle.

ANNEXE 1

Corpus d'exclusion lexicographique

1. Pour la langue catalane:

Enciclopèdia catalana (1982). *Diccionari de la llengua catalana.*

Enciclopèdia catalana (1993). *Diccionari de la llengua catalana.*

Enciclopèdia catalana (1990). *Gran enciclopèdia catalana* (2a. edició).

FABRA, P. (1932). *Diccionari general de la llengua catalana.*

2. Pour la langue castillane:

Real Academia Española (1992). *Diccionario de la lengua española.*

Real Academia Española (1989). *Diccionario manual e ilustrado de la lengua española.*

VOX-Biblograf (1987). *Diccionario general ilustrado de la lengua española.*

ANNEXE 2

Observatori de neologia Fitxa de buidatge Data:

1

antidisturbis

2 m pl | 3 | 4 DLC GEC2

5

Com era de perveure, van venir els *antidisturbis* al cap de poc i van començar a repartir llenya a tort i dret...

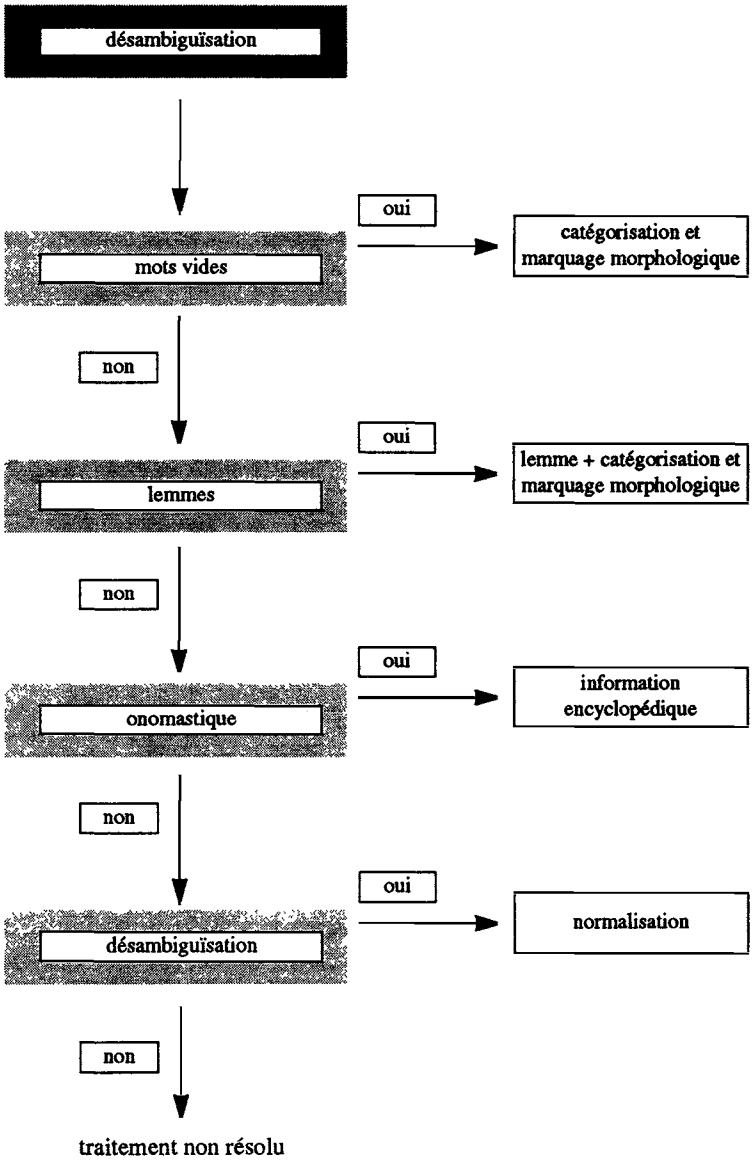
6

7 DB | a 14-10-90 | b POLIT | c 35 | d | 8 TRA 01-12-90 | 9 ASS | 10 18-12-90

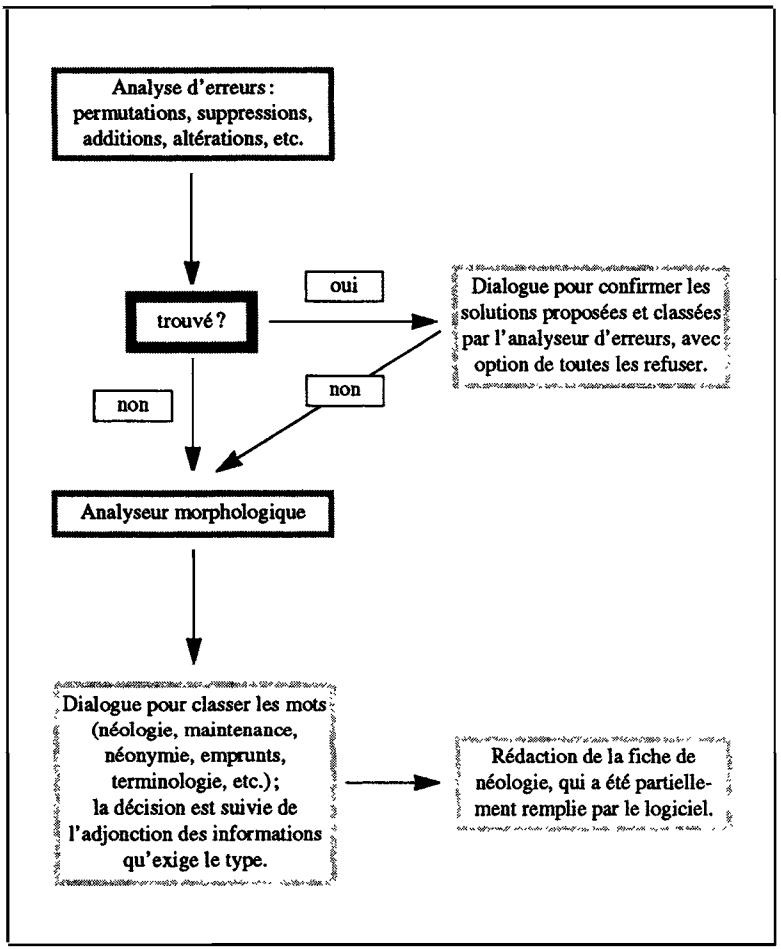
Identification des champs dans la fiche:

Numéro	**Code**	**Signifié**
1	EN	Entrée
2	CG	Catégorie grammaticale
3	AT	Aspects typographiques
4	DC	Dictionnaires consultés (corpus d'exclusion)
5	CO	Contexte
6	NO	Note
7	RF	Référence: a. Source b. Date c. Section d. Page
8	AU/DT	Auteur et date de la fiche
9	AV	Auteur de la vérification
10	DV	Date de la vérification

ANNEXE 3


désambiguïsation
oui
mots vides
catégorisation et marquage morphologique
non
oui
lemmes
lemme + catégorisation et marquage morphologique
non
oui
onomastique
information encyclopédique
non
oui
désambiguïsation
normalisation
non
traitement non résolu

ANNEXE 4


Analyse d'erreurs : permutations, suppressions, additions, altérations, etc.
trouvé ?
oui
non
non
Dialogue pour confirmer les solutions proposées et classées par l'analyseur d'erreurs, avec option de toutes les refuser.
Analyseur morphologique
Dialogue pour classer les mots (néologie, maintenance, néonymie, emprunts, terminologie, etc.); la décision est suivie de l'adjonction des informations qu'exige le type.
Rédaction de la fiche de néologie, qui a été partiellement remplie par le logiciel.

M. Teresa Cabré [Courrélec.: cabre@upf.es] et Lluis de Yzaguirre [Courrélec: de_yza@upf.es]: Institut de Lingüística Aplicada, Universitat Pompeu Fabra, Balmes, 132 – 08008 Barcelona España

RÉSUMÉ: Stratégie pour la détection semi-automatique des néologismes de presse – Les auteurs présentent l'Observatori de Neologia de Barcelona (OBNEB). L'un des objectifs de l'organisme est la détection de la néologie formelle dans la presse écrite. L'article expose les concepts de base qui sont en rapport avec la néologie. Puis il décrit les méthodes de travail de l'OBNEB. Ensuite, les caractéristiques du corpus sont détaillées. Ce corpus est riche de sept millions d'occurrences provenant du dépouillement de textes journalistiques. Les premiers essais d'extraction semi-automatique des néologismes ont été menés à partir de ces données. Enfin, les auteurs expliquent le fonctionnement du logiciel qui est utilisé pour repérer et pour traiter les néologismes. Ce logiciel a été mis au point à l'Institut universitari de lingüística aplicada de la Universitat Pompeu Fabra. D'autres informations sur les travaux des auteurs peuvent être consultées à l'adresse suivante: «http://www.iula.upf.es».

ABSTRACT: A Strategy for the Semi-Automatic Detection of Press Neologisms – The Observatori de Neologia de Barcelona (OBNEB) was created in order to automatize the detection of formal neologisms found in the press. In this article the basic concepts related to neology are introduced together with the process carried out at the OBNEB. The features of the Corpus of the University of Barcelona are presented in detail. Composed of seven millions of occurrences taken from press texts, this corpus was used to test a programme for the semi-automatic extraction of neologisms. Finally, the authors show how the programme developed at the *Institut de Lingüística Aplicada* manages to fulfill this task. Further information about the authors and their work can be found in "http://www.iula.upf.cs".