 

# CONTRIBUTION À L'ANALYSE ET À LA GÉNÉRATION AUTOMATIQUES DE NÉOLOGISMES À L'AIDE DE GRAPHES CONCEPTUELS

**Henri ZINGLÉ**

*LILLA*
*Université de Nice-Sophia Antipolis*
*BP-209 F-06204 Nice Cedex*
*henri@lilla.unice.fr*

Résumé : On propose un système d'analyse et de génération automatiques de néologismes utilisant une base de connaissances dans laquelle sont explicitées les contraintes formelles et sémantiques des formants ainsi qu'une méthode de représentation du sens fondée sur la théorie des graphes conceptuels. L'approche utilisée s'applique à la fois aux dérivés et aux composés; elle est adaptable à différentes langues à condition de construire les bases de connaissances requises. On insiste sur les problèmes de représentation du sens et les principes de formalisation avant d'aborder les aspects pratiques à proprement parler. Les applicatifs développés sont intégrés à l'environnement de génie linguistique ZStation et conduisent à des applications concrètes en veille terminologique et en aménagement linguistique.

Mots clés : néologie - traitement automatique des langues - graphes conceptuels

Dans le cadre d'une étude sur les néologismes on a développé un ensemble de processus permettant de représenter le sens de mots dérivés et composés et de façon symétrique de les générer à partir d'une représentation sémantique. L'étude a été limitée aux cas de dérivation et de composition ***calculables***, en privilégiant une approche cognitive fondée sur le sémantisme des unités morphologiques (UMs) et le sémantismes des relations entre concepts. La démarche adoptée unit dérivation et composition, en introduisant comme seuls critères

ISBN: 0 08 043 438X

 

typologiques la capacité des UMs à admettre une expansion à gauche ou à droite et à devenir ou non gouverneurs de construction. On présentera ici succinctement les choix effectués concernant la méthode de représentation du sens, la formalisation des connaissances linguistiques et l'implantation informatique des processus d'analyse et de génération.

## 1. REPRÉSENTATION DU SENS

Pour la représentation du sens on prend appui sur le formalisme des graphes conceptuels (GCs) de Sowa (1984) qui permet d'exprimer les caractéristiques intrinsèques des concepts ainsi que les relations qui les unissent, à des niveaux de complexité variable. Un GC canonique se compose de deux concepts unis par une relation [Cpt1]→(Rel)→[Cpt2] (ici en notation non graphique) ; par principe le nombre de relations est réduit au minimum.

La représentation du sens de mots dérivés et composés à l'aide de ce formalisme réclame cependant un aménagement du formalisme. En effet, si l'on représente *enregistreur* par [enregistrer]→(instr)→[appareil], rien ne permet de dire si l'on exprime le fait qu'un appareil est entrain d'enregistrer ou si l'on désigne l'appareil lui-même et la fonctionnalité qui lui est associée. Ce problème peut toutefois être résolu en jouant sur le fait qu'un concept peut être noté soit [chat] soit [animal:chat], cette dernière notation faisant apparaître qu'un chat est une instance d'animal. A l'instar de la notation proposée par Sowa lui-même où le types de PROPOSITION et de SITUATION admettent un GC comme instance, on propose d'étendre cette propriété à l'ensemble des types. Ainsi le sens du dérivé *enregistreur* serait représenté par le graphe [appareil:[enregistrer]→(instr)→[appareil]] qui peut être mis en relation avec l'énoncé *appareil qui est l'instrument utilisé pour enregistrer* ; de façon analogue, le composé *boîte à chocolats* peut être représenté par [boîte:[boîte]→(cont)→[chocolats]]. Bien entendu, le principe peut être appliqué de façon récursive : [dictionnaire:[dictionnaire]→(cont)→[termes:[termes:[termes]→(chrc)[officiels]]] pour fr. *dictionnaire des termes officiels* ou [handle:[handle]→(loc)→[tür:[tür]→(loc)→[haus]]] alld. *Haustürgriff*.

Afin de rendre plus cohérent le traitement de la modalité - assez flou au demeurant chez Sowa - on propose également d'introduire la relation **var** qui évite d'avoir recours à des relations unaires ([Cpt]→(var)→[possible] au lieu de (psbl)→[Cpt]).

## 2. PRINCIPES DE FORMALISATION

L'analyse et la génération automatiques de néologismes à l'aide des GCs utilisent les mêmes connaissances, formalisées sous forme déclarative en accord avec l'approche générale qui sous-tend la ZStation (Zinglé, 1994). Celles-ci se répartissent en deux bases de connaissances indépendantes centrées respectivement sur les propriétés morphologiques (intralinguistiques) et les propriétés conceptuelles (interlinguistiques).

Les entrées de la base de connaissances morphologiques sont soit des morphes (ou des combinaisons de morphes) soit des unités lexicales simples. Pour chacune d'elles on indique : (a) sa capacité d'être gouverneur de construction (b) si l'argument éventuel s'insère à gauche ou à droite du gouverneur (c) la catégorie et les variables morphosyntaxiques de la construction produite (d) le GC virtuel associé (e) et un ensemble de contraintes formelles sémantiques. La codification permet de traiter des situations très divergentes (cf. fr. *chambre de commerce* vs alld. *Handelskammer*).

ISBN: 0 08 043 438X

 

Les propriétés conceptuelles sont formalisées sous forme de réseaux multidimensionnels à héritage multiple que l'on peut aisément constituer avec la ZStation. Pour mémoire, chaque noeud du réseau est caractérisé par un ensemble de relations conceptuelles, ce qui permet de connaître à tout moment l'ensemble des propriétés d'un concept donné ou de vérifier s'il est en mesure de satisfaire une contrainte sémantique particulière. Pour des raisons de lisibilité et de portabilité il a été décidé de choisir des identificateurs de concept en anglais suivis d'un indice numérique (par exemple, water0 pour le concept d'eau water1 pour le concept arroser).

## 3. ANALYSE ET GÉNÉRATION AUTOMATIQUES

Les processus d'analyse et de génération automatiques ont été implantés en PDC-Prolog (version 4.1 sous Windows). Ils sont intégrés à l'environnement de génie linguistique ZStation. Ils sont fondamentalement indépendants des langues naturelles, les stratégies de traitement propres à chaque langue étant codifiées dans la base de données morphologiques. Toutes dispositions utiles ont été prises pour favoriser l'interactivité et permettre à l'ingénieur linguiste de passer aisément de la formalisation des données linguistiques à l'analyse ou à la génération.

L'analyse d'un mot composé ou dérivé fait appel à un générateur d'hypothèses dont le rôle est de construire une hiérarchisation des UMs détectées. Pour chaque construction impliquant deux UMs on vérifie, si les contraintes formelles et sémantiques sont satisfaites, en fonction du regroupement à gauche ou à droite. En cas de succès on détermine le statut de l'unité intégrante et on procède à l'instanciation de l'argument dans le GC virtuel qui lui est associé. La complexité du traitement lorsqu'il y a plus de deux unités en présence est traité en élaguant l'arbre des solutions au moment même de sa construction. L'utilisation des contraintes sémantiques permet, entre autres, de calculer des solutions différentes pour *arroseur* ([person0:[water1]→(agnt)→[ person0]] et [device0:[water1]→(instr)→[ device0]]) opposé à *camionneur* ([person0:[person0]←(agnt)←[drive0]→(obj)→[truck0]]) et pour *tasse de café* ([tea0:[coffee]→(loc)→[cup0]]) opposé à *tasse de porcelaine* ([cup0:[cup0]→(matr)→ [porcelain0]]).

Le processus de génération prend en compte (a) une représentation sémantique sous forme de GCs et (b) la sélection du statut morpho-syntaxique de l'unité à produire. On peut ainsi générer aisément à partir du même graphe [possible0:[record0]→(var)→[possible0]] l'adjectif *enregistrable* et le substantif *enregistrabilité*. De la même façon [false0:[possible0: [record0]→(var)→[possible0]→(var)→[false0]] permet de générer *inenregistrable* et *inenregistrabilité*.

## PERSPECTIVES

Outre la modélisation linguistique et ses applications dans l'enseignement de la lexicologie, les processus développés ouvrent des perspectives intéressantes en aménagement terminologique, soit pour automatiser la veille terminologique soit pour forger de toutes pièces des néologismes officiels. On notera en passant que l'approche proposée permet éventuellement de créer un néologisme dans une langue cible à partir de l'analyse d'un terme en langue source et rejoint ainsi une préoccupation majeure partagée par de nombreux spécialistes de l'aménagement terminologique.

ISBN: 0 08 043 438X

 

RÉFÉRENCES

Sowa, J. (1984) *Conceptual structures : information processing in man and machine.* Addison-Wesley.

Zinglé, H. (1994) The ZStation workbench and the modelling of linguistic knowledge. *Current issues in mathematical linguistics*, Elsevier-NHLS, 1994, 423-432

Zinglé, H. (1996) Analyse et génération automatiques de mots dérivés en français, *Travaux du LILLA*, Université de Nice-Sophia Antipolis pp 27-42

ISBN: 0 08 043 438X