Paradigmes de flexion et traits syntactico-sémantiques dans le dictionnaire électronique de catalan

Judith Sastre Alaiz
Universitat Autònoma de Barcelona

Dans le cadre du système de dictionnaires électroniques du Groupe *f*LexSem (Phonétique, Lexicologie et Sémantique) de l'Université Autonome de Barcelone, nous avons élaboré un prototype de dictionnaire électronique du catalan coordonné avec l'espagnol et le français.

Le dictionnaire comporte, à l'heure actuelle, plus de 35 000 lemmes qui ont été puisés de textes catalans procédant de différentes sources en langue standard, notamment, d'un corpus de près de 5 Mo de texte provenant de pages Web en catalan.

La microstructure du dictionnaire a été développée d'après le format communément utilisé dans les dictionnaires du groupe *f*LexSem, qui contemple des champs morphologiques, syntaxiques, sémantiques, diasystématiques et de traduction. Le but de ce prototype étant d'obtenir un premier dictionnaire électronique du catalan le plus complet possible du point de vue de la reconnaissance de formes simples, nous nous sommes centrée, dans un premier moment, sur la description de la morphologie flexionnelle.

Dans les dictionnaires en papier ou les dictionnaires en support électronique issus directement de dictionnaires conçus en format papier, les informations sur la morphologie flexionnelle ne sont données que pour certaines entrées dont le choix est extrêmement arbitraire. Un dictionnaire électronique ne peut pas se permettre un tel procédé, toute information doit être explicitement indiquée pour toutes et chacune des entrées. D'une perspective de génération de texte (synthèse) ou d'une perspective de reconnaissance (analyse), il est indispensable que le dictionnaire électronique dispose des informations nécessaires pour que le système puisse générer et/ou reconnaître toutes les formes d'un mot. À cet effet, nous avons construit une collection d'automates à états finis couvrant la totalité des paradigmes flexionnels pour les noms, les adjectifs et les verbes du catalan. Les automates nous ont permis de générer automatiquement toutes les formes fléchies correspondantes

aux entrées de notre dictionnaire sans avoir besoin de connaître un langage de programmation spécifique. Pour ce faire, nous avons travailler avec le logiciel Intex (© 2005 Max Silberztein, Université de Franche-Comté) ainsi qu'avec sa nouvelle version NooJ (© 2005 Max Silberztein, Université de Franche-Comté). Par le biais de l'éditeur de graphes incorporé à ces logiciels, nous avons réalisé les modèles de flexion qui décrivent les différentes opérations à effectuer pour générer les formes fléchies. Ainsi, chaque entrée du dictionnaire est associée par un code au modèle de flexion qui lui correspond. Nous disposons d'une collection de 193 graphes avec NooJ et de 233 graphes avec Intex répartis comme suit:

INTEX	***NOOJ***
39 graphes pour les substantifs	29 graphes pour les substantifs
64 graphes pour les adjectifs	50 graphes pour les adjectifs
131 graphes pour les verbes	114 graphes pour les verbes

La différence du nombre de graphes entre une version et une autre est due aux différents opérateurs intégrés dans l'éditeur de graphes. NooJ à différence d'Intex dispose d'un opérateur qui peut modifier le type d'accent des mots sans qu'il n'y ait des répercussions sur la voyelle. Pour la description formelle de la flexion d'une langue à accent libre comme le catalan où des modifications diverses peuvent se produire au niveau de l'accent telles que l'apparition, la disparition ou le changement de type d'accent, cet opérateur est de grande utilité, car un seul modèle de flexion est valable pour les cinq voyelles. Par exemple:

masc. sing.	***masc. pl.***	***fém. sing.***	***fém. plu.***
escàs	escasos	escasa	escases
francès	francesos	francesa	franceses
indecís	indecisos	indecisa	indecises
escrupolós	escrupolosos	escrupolosa	escrupoloses
difús	difusos	difusa	difuses

Avec NooJ ces cinq mots correspondent à un seul modèle tandis qu'avec Intex nous avons dû créer cinq graphes différents, un pour chaque voyelle.

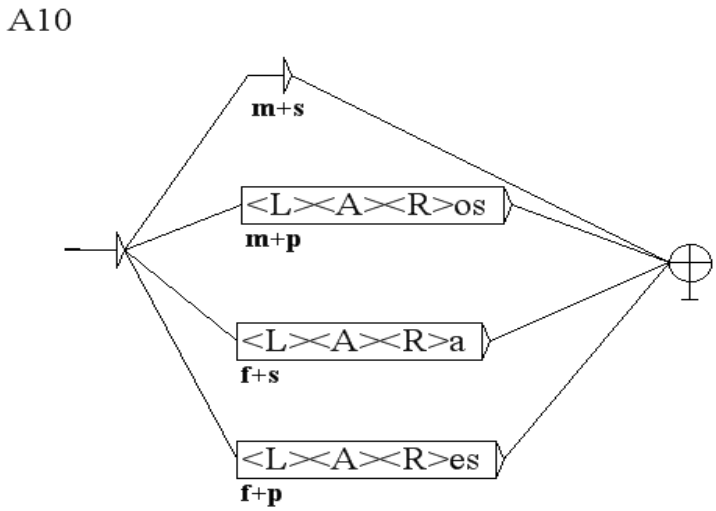

Figure 1.1: Exemple de graphe de flexion avec NooJ

Le système de codage adopté nous donne à simple vue diverses informations. La première lettre nous indique la catégorie grammaticale (Adjectif, Nom ou Verbe) et dans le cas des substantifs le numéro qui suit nous indique le genre. Si le nombre est compris entre 01 et 49 il s'agit d'un nom masculin; si le nombre est compris entre 50 et 99 il s'agit d'un nom féminin. Remarquons que nous n'avons pas traité le genre naturel dans le cadre de la morphologie flexionnelle car nous le considérons un phénomène dérivationnel.

Lorsque nous compilons notre dictionnaire électronique du catalan avec NooJ ou avec Intex, de 35 870 entrées le programme de flexion génère un total de 448 900 entrées. Ces chiffres concordent avec les études réalisées par IBM sur la langue espagnole qui montrent qu'un dictionnaire de 35 000 mots simples donnera plus de 400 000 mots fléchis.

Mis à part les informations sur la morphologie flexionnelle, dans le cas des substantifs nous avons fait un premier classement en traits syntactico-sémantiques: *humain, animal, végétal, concret, abstrait, locatif et temporel*. Ce classement nous a conduit à réaliser les premiers dédoublements sémantiques au sein du dictionnaire; le but étant que chaque entrée du dictionnaire constitue une unité lexicale, c'est à dire un triplet constitué d'une forme (*a)*, d'un sens ('a') et d'une combinatoire (∑a) (Mel'čuk, 1995). Ainsi, par exemple le mot *flamenc* correspond à deux unités lexicales dans notre dictionnaire, i.e deux entrées différentes. L'une est accompagnée de l'étiquette *Anl,* l'autre de l'étiquette *Abst.* Pour certains arguments, nous avons

complété cette description à l'aide des classes d'objets (Gross 1994; Le Pesant et Mathieu-Colas 1998). Il s'agit d'un sous-classement à l'intérieur des traits, délimitant encore plus la nature syntactico-sémantique des entrées. Chaque classe d'objets est définie par ses prédicats appropriés. Ainsi, par exemple, la classe d'objets <oiseau> a comme prédicats verbaux appropriés: *plomar; bequetejar-se; aletejar; covar*. La caractérisation en traits syntactico-sémantiques et classes d'objets constitue un moyen efficace à la levée d'ambiguïtés. Dans les exemples ci-dessous, chaque forme verbale est une unité lexicale différente.

plomar/N0:Hum/N1:Conc/Fr:plomber

plomar/N0:Hum/N1:Anl<oiseau>/Fr:plumer

plomar/N0:Anl<oiseau>/Fr: se couvrir de plumes

afaitar/N0:Hum/N1:Hum/Fr:raser

afaitar/N0:Hum/N1:Anl<oiseau>/Fr:dresser

aletejar/N0:Anl<oiseau>/Fr:battre des ailes

aletejar/N0:Anl<poisson>/Fr:agiter les nageoires

Notre dictionnaire dispose également d'un champ destiné à la traduction. Chacune des entrées devra être accompagnée d'un équivalent de traduction vers le français et vers l'espagnol. Actuellement nous avons établi près de 3 000 équivalents de traduction.

Dans un futur immédiat, nos travaux suivront une double voie: d’une part, nous compléterons la macrostructure du dictionnaire moyennant le dépouillement systématique de textes et nous entreprendrons la prise en compte de différentes variantes du catalan. D’autre part, nous compléterons la microstructure avec l’introduction des structures d’arguments pour les prédicats et par rapport à la version plurilingue du dictionnaire avec l'introduction des équivalents de traductions pour chaque entrée.

Références bibliographiques

BOUILLON, P. (1998). *Traitement automatique des langues naturelles*. Champs Linguistiques. Duculot, Belgique.

BLANCO, X. (1999). *Lexicographie bilingue français-espagnol et classes d'objets*. Col·lecció Materials (73). Servei de publicacions de l'UAB, Barcelona.

BLANCO X. (1997) «Noms composés et traduction français-espagnol», *Lingvisticæ Investigationes* XXI:1, John Benjamins B.V., Amsterdam.

BLANCO, X. (2001) «Dictionnaires électroniques et traduction automatique espagnol-français» *Langages, 143,* p. 49-70*,* Larousse, Paris.

COURTOIS B. et SILBERZTEIN M. (1990) «Dictionnaires électroniques du français», *Langue Française* 87, Larousse, Paris.

GROSS G. (1990). «Définition des noms composés dans un lexique-grammaire», *Langue Française* 87, Larousse, Paris.

GROSS, G. (1992). «Forme d'un dictionnaire électronique». In *L'environnement traductionnel actes du colloque de Mons*, pages 255–271,Mons. AUPELFUREF.

GROSS G., (1994). «Classes d'objets et description des verbes», Langages 115, Larousse, Paris.

GROSS,M. (1998). «La fonction sémantique des verbes supports». *Travaux de Linguistique*, (37): 25–46. Bruxelles, Duculot.

LAMIROY B. (1998). «Le lexique-grammaire. Essai de synthèse», *Travaux de Linguistique.* (37), Bruxelles, Duculot.

LE PESANT D. et MATHIEU-COLAS, M. (1998). «Introduction aux classes d'objets», *Langages* 131, Paris, Larousse.

MATHIEU-COLAS, M. (1994). *Les mots à traits-d'union.* Didier Erudition, Paris.

MEL'ČUK, I. (1984). *Dictionnaire explicatif et combinatoire du français contemporain* Recherches lexico sémantiques I. Les Presses de l'Université de Montréal, Montréal.

MEL'ČUK, I. (1993). *Cours de morphologie générale,* volume 1. Les Presses de l'Université de Montréal CNRS, Montréal.

MEL'ČUK, I. (1994). «Cours de morphologie générale» volume 2. Les Presses de l'Université de Montréal CNRS, Montréal.

MEL'ČUK, I., Clas, A. & POLGUÈRE, A. (1995). *Introduction à la lexicologie explicative et combinatoire*. Champs linguistiques. Duculot.

SILBERZTEIN, M. (1990). «Le dictionnaire électronique des mots composés». *Langue Française*, (87): 71–83.

SILBERZTEIN, M. (1993). *Dictionnaires électroniques et analyse automatique de corpus: Le système INTEX*. Masson, Paris.