TALN 24-26 Juin 2009

Intégrer les tables du Lexique-Grammaire à un analyseur syntaxique à grande échelle

Benoît Sagot[1], Elsa Tolone[2]

1. Alpage, INRIA & Université Paris 7
Domaine de Voluceau, Rocquencourt, BP 105, 78153 Le Chesnay
benoit.sagot@inria.fr

2. IGM, Université Paris-Est
5 bd Descartes, Champs-sur-Marne, 77454 Marne-la-Vallée
elsa.tolone@univ-paris-est.fr

RESUME

Dans cet article, nous montrons comment nous avons converti les tables du Lexique-Grammaire en un format TAL, celui du lexique Le*fff*, permettant ainsi son intégration dans l'analyseur syntaxique FRMG. Nous présentons les fondements linguistiques de ce processus de conversion et le lexique obtenu. Nous validons le lexique obtenu en évaluant l'analyseur syntaxique FRMG sur le corpus de référence de la campagne EASy selon qu'il utilise les entrées verbales du Le*fff* ou celles des tables des verbes du Lexique-Grammaire ainsi converties.

ARCHITECTURE

Le Lexique-Grammaire

Lexique syntaxique organisé par classes (Gross 75) :
- Une *table des classes* par catégorie rassemble les informations vraies sur des classes entières
- Une *table* par classe rassemble les informations qui varient d'une entrée à l'autre
- Les entrées sont distinguées sémantiquement

V_13

N0 =: Nhum	N0 =: Nnr	N0 =: le fait Qu P	N0 =: V1-inf W	N0 =: V2-inf W	Ppv	<ENT>	N0 V	N0 V N1	N1 =: Nhum	N1 =: N-hum	N0 V N2	N1 =: Qu Psubj	N0 être V-ant pour N1	N1 se V de ce Qu P	N1 se V auprès de N3 de ce	
+	+	+	+	-	<E>	soulager	+	+	+	-	+	-	+	+	+	+

Table des classes	V_13	V_14
N0 V N0pc W	?	?
N0 V N1	o	?
N0 V N1 à ce Qu P	?	?
N0 V N1 à N2	-	-
N0 V N1 à N2hum	?	?
N0 V N1 de N2	+	?
N0 V N1 de N2 source	?	?

lglex

Format Alexina et lexique Le*fff*

Alexina est une architecture pour les lexiques morphologiques et syntaxiques, et notamment le lexique Le*fff* du français (Sagot *et al.* 06)
C'est le format des lexiques utilisés par l'analyseur syntaxique FRMG (Thomasset & de la Clergerie 05)

ID = V_13_144
lexical-info=[cat="verb", verb=[lemma="soulager"]]
args=(**const**=[pos="0", dist=(comp=[cat="NP", nothum="true"], comp=[cat="NP", hum="true"], comp=[cat="comp", mood="ind"], comp=[cat="comp", mood="subj"], comp=[cat="inf"], comp=[cat="leFaitComp"], comp=[cat="inf", contr="1"])],
const=[pos="1", dist=(comp=[cat="NP", hum="true"])],
const=[pos="2", dist=(comp=[cat="NP", nothum="true"], comp=[cat="ceComp", mood="subj"], comp=[cat="inf", contr="1"])])
all-constructions=[**absolute**=(construction="o::N1 être Vpp par N0", construction="o::N0 être V-ant pour N1", construction="true::N0 V N1 de N2", construction="o::N0 V", construction="o::N1 être Vpp par N0 de ce Qu P", construction="o::N1 se V de ce Qu P", construction="o::N0 V N1", construction="o::N0 V N2", construction="o::N1 se V auprès de N3 de ce Qu P"),
relative=(construction="Prép N2 =: Prép Qu P = Ppv")]
example=[example="Max a soulagé Ida de ce qu'elle ait à tout faire"]

Conversion au format Alexina

Objectif : utiliser les tables des verbes simples du Lexique-Grammaire dans FRMG
Conversion des tables au format *lglex* (Constant & Tolone 08) vers le format Alexina :
- découpage en entrées (constr. de base + variantes)
- construction des cadres de sous-catégorisation
- identification des redistributions admissibles
- informations complémentaires

1 Entrées	2 Cadres de sous-catégorisation	3 Redistributions
Construction de base N0 V N1 de N2 + variantes N0 V et N0 V N1	<**Suj**:cln|scompl|sinf|sn**, Obj**:(sn)**,** **Objde**:(de-scompl|de-sinf|de-sn|en)>	%actif,%passif, %ppp_employé_comme_adj
Entrée secondaire N0 V N2 (redistribution non standard)	<**Suj**:cln|scompl|sinf|sn**, Obj**:sn|cla>	%actif,%passif, %ppp_employé_comme_adj
Entrée secondaire N1 se V auprès de N3 de ce Qu P (emploi distinct)	<**Suj**:cln|scompl|sinf|sn**, Obl**:auprès_de-sn**,** **Objde**:de-scompl|de-sinf|de-sn|en>	%actif
Entrée secondaire N1 se V de ce Qu P (emploi distinct)	<**Suj**:cln|scompl|sinf|sn**, Objde**:de-scompl|de-sn>	%actif

soulager___13_144 v-er:std — **Entrées finales au format Alexina** (4)

100;**Lemma**;v;<**Suj**:cln|scompl|sinf|sn**, Obj**:(sn)**, Objde**:(de-scompl|de-sinf|de-sn|en)>; cat=v,@DeCompSubj, @CtrlObjSuj,@CtrlObjObjde; %actif,%passif,%ppp_employé_comme_adj
construction de base N0 V N1 de N2 (N0 V ; N0 V N1)

100;**Lemma**;v;<**Suj**:cln|scompl|sinf|sn**, Obj**:sn|cla>; cat=v,@CompSubj,@CtrlSuj,@CtrlObj; %actif,%passif, %ppp_employé_comme_adj # N0 V N2

100;**se Lemma**;v;<**Suj**:cln|scompl|sinf|sn**, Obl**:auprès_de-sn**, Objde**:de-scompl|de-sinf|de-sn|en>; cat=v,@DeCompSubj,@CtrlSuj,@CtrlObjde; %actif # N1 se V auprès de N3 de ce Qu P

100;**se Lemma**;v;<**Suj**:cln|scompl|sinf|sn**, Objde**:de-scompl|de-sn>;cat=v,@CtrlObjdeSuj;%actif # N1 se V de ce Qu P

Résultat

Nous obtenons ainsi une représentation au format Alexina de la plupart des informations codées dans les tables de verbes du Lexique-Grammaire
- 16 903 entrées verbales (Le*fff* : 7 072) décrivant 5 694 lemmes (Le*fff* : 6 818)
- Entrées verbales complétées par les entrées du Le*fff* pour les autres catégories, les auxiliaires…

ÉVALUATION

Évaluation

Évaluation sur le corpus EASy de l'analyseur syntaxique FRMG avec le lexique Le*fff* et avec notre lexique **préliminaire** issu de *lglex*
- Rappel: *lglex* est sans rapport avec les choix EASy
- $FRMG_{Lefff}$ reste meilleur pour l'instant (f-mesure)
- $FRMG_{lglex}$, dépasse parfois $FRMG_{Lefff}$, tantôt sur les chunks, tantôt sur les relations
- Les taux de couverture sont similaires, mais les phrases couvertes sont significativement différentes

	Chunks EASy		Relations EASy	
	Le*fff*	*lglex*	Le*fff*	*lglex*
lemonde	**86,8%**	82,8%	**59,8%**	56,9%
senat	82,7%	**83,1%**	**56,7%**	54,9%
oral	**74,1%**	73,6%	47,2%	**48,5%**
TOTAL	**84,4%**	82,3%	**59,9%**	56,6%

Perspectives

- Poursuite du développement de la table des classes (p. ex. sur le passif)
- Recherche automatique d'erreurs par utilisation à grande échelle ($FRMG_{lglex}$)
- Amélioration du convertisseur *lglex* → Alexina présenté ici
- Utilisation conjointe de $FRMG_{Lefff}$ et $FRMG_{lglex}$, en favorisant un système s'il a pu construire une analyse complète