Les tables du Lexique-Grammaire au format TAL

Elsa Tolone[1]

1. IGM, Université Paris-Est (France)

Conférence MajecSTIC – Avignon, France
16 novembre 2009

Contexte

- Les **tables du Lexique-Grammaire** sont une source très riche d'informations lexicales
- Ces informations devraient être très utiles pour l'**analyse syntaxique** automatique
- Mais les tables ne sont **pas exploitables** telles quelles dans un analyseur syntaxique
  - des informations importantes ne sont pas représentées car elles sont considérées comme implicites pour une table donnée
  - le format de représentation des informations lexicales n'est pas formalisé
  - il faut interfacer ces informations avec un analyseur syntaxique réel

1. L'analyse syntaxique

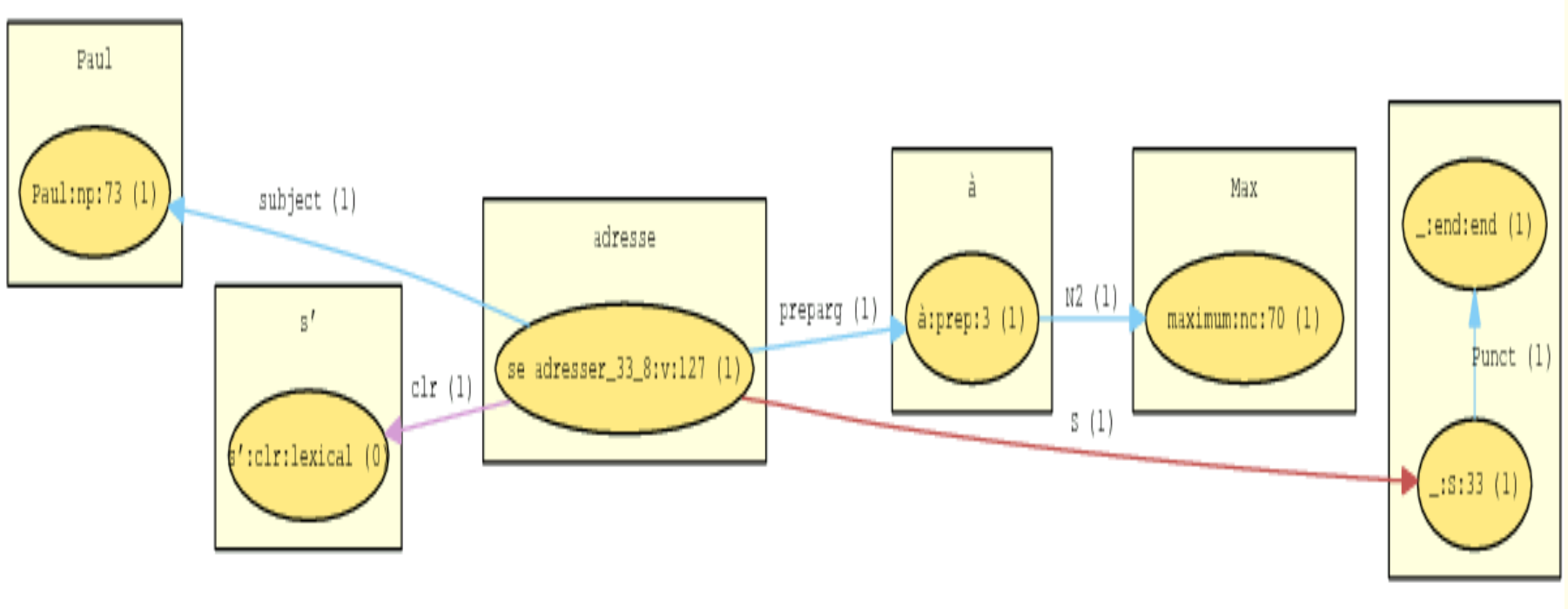
Analyse syntaxique

- **Construire la structure grammaticale d'une phrase**
  pour lui donner du sens
  $\longrightarrow$ en explicitant les relations de dépendance entre les mots (entre sujet et objet par ex.)
  Difficulté : **Complexité et grandeur de la langue**
- Applications :
  - Compréhension de texte
  - Extraction d'information
  - Traduction

Analyseurs syntaxiques

- **symboliques** = grammaire + lexique syntaxique qui spécifie le comportement grammatical de chaque mot de la langue
  $\longrightarrow$ développés manuellement
- **probabilistes** = modèle acquis à partir d'un corpus annoté manuellement
- Analyseur syntaxique symbolique retenu :
  FRMG [Thomasset & de La Clergerie 2005]

Exemple de dépendances



Paul s'adresse à Max

2 Les tables du Lexique-Grammaire

Tables du Lexique-Grammaire

- Caractéristiques :
  - Développées **manuellement** depuis plus de 30 ans par l'Equipe d'Informatique Linguistique de l'IGM (Université Paris-Est)
  - Décrivent caractéristiques grammaticales des mots
  - Ressources lexicales riches
  - Pas utilisables directement dans un analyseur symbolique
- Objectifs :
  - **adapter et convertir** les tables du Lexique-Grammaire en un format TAL
  - **coupler** le lexique syntaxique obtenu, nommé $lglex_{\mathrm{Lefff}}$, avec un analyseur syntaxique
  - **évaluer** l'analyseur obtenu

Historique

- [Gross 1975]
- Étude de la syntaxe d'une phrase élémentaire
  (ou **cadre de sous-catégorisation**)
  ex : $N_0\ V\ N_1$
- Usages des verbes, adverbes, noms et adjectifs prédicatifs et expressions figées pour le français
  $\longrightarrow$ certaines **propriétés en commun**
- Les différents emplois sont distingués (ex : *cuisiner*)
- 13,400 verbes simples dans **61 classes**

Principe

- Chaque classe est codée dans une **table** :
  - lignes = entrées (mots)
  - colonnes = propriétés grammaticales
- **Propriétés** :
  - cadres de sous-catégorisation
  - autres propriétés (distributionnelles, morphologiques, transformationnelles, sémantiques,. . .)
    ex : N_0 := *Nhum* $\rightarrow$ noms de personnes
- Chaque propriété **testée pour chaque entrée**
  $\rightarrow$ codage binaire :
  - $+$: propriété acceptée
  - $-$: propriété non acceptée
- **Propriétés définitoires** pour chaque classe

Extrait de la table des verbes de la classe 33

N0 =: Nhum	N0 =: N-hum	N0 =: Nmr	Ppv	Ppv =: se figé	Ppv =: en figé	Ppv =: les figé	Nég	<ENT>	N0 V	N0 être V-ant	N1 =: Nhum	N1 =: N-hum	N1 =: le fait Qu P	Ppv =: lui	Ppv =: y	N0hum V W sur ce point	[extrap]	<OPT>
+	-	-	<E>	-	-	-	-	renaître	+	+	-	+	-	-	+	-	-	Max renaît au bonheur de vivre
+	-	-	se	+	-	-	-	rendre	+	-	+	+	+	-	+	+	+	Max s'est rendu à mon opinion
+	-	-	se	+	-	-	-	rendre	+	-	+	-	-	-	-	-	-	Le caporal s'est rendu à l'ennemi
+	-	-	<E>	-	-	-	-	renoncer	-	-	+	+	-	-	+	-	-	Max renonce à son héritage

Propriété définitoire : N_0 V à N_1

Bilan

- Inventaire :
  - 61 classes de **verbes simples**
  - 32 classes d'**adverbes** (adverbes en *-ment* et locutions adverbiales)
  - 59 classes de **noms prédicatifs** (noms avec argument(s) qui sont étudiés avec leur verbes support)
    ex : *Luc monte une attaque contre le fort*
  - 65 classes d'**expressions figées**
    ex : *arriver à la cheville*
- Avantages :
  - Description riche
  - Large couverture
  - Base linguistique solide

Problèmes

- **Noms différents** pour une même propriété
  → Harmonisation des intitulés de colonnes
  ex : *[extrap]* et *il V N_0 W*
- Propriétés **pas définies clairement**
  → Documentation des propriétés
- **Propriétés définitoires implicites** (littérature)
  → Constante + ou − pour l'ensemble de la table
- Toutes les propriétés ne sont **pas codées** dans chaque table
  → Codage +, − ou *o* pour l'ensemble de la table

Travail en cours pour les verbes, adverbes, noms prédicatifs, et expressions figées

Tables des classes

Propriétés définitoires → à décrire dans des **tables des classes** (1 par catégorie) :

- colonnes = toutes les propriétés syntaxiques répertoriées pour la catégorie concernée (après normalisation)
- lignes = classes définies pour cette catégorie
- intersection ligne/colonne :
  - o = propriété explicitement codée dans la table concernée
  - $+$ (resp. $-$) = la propriété correspondante est vérifiée (resp. non vérifiée) par tous les éléments de la classe
  - ? = non encore renseigné

En cours de réalisation à l'IGM, presque terminé pour les verbes et les noms prédicatifs [Constant & Tolone 2008]

Extrait de la tables des classes des verbes

table	N0 =: Nhum	N0 =: N-hum	N0 =: Nnr	N0 =: V1-inf W	<ENT>	Ppv =: se figé	N0 V	N0 V N1	zone 1	N0 V à N1	N1 =: Nhum	N1 =: N-hum	N0 V Prep N1 V0-inf W	N0 V N1 V0-inf W	N0 V V0-inf W
V_2	+	-	-	-	o	o	-	-	-	-	-	+	o	o	+
V_4	-	-	+	+	o	-	o	+	-	-	o	o	-	-	-
V_31R	o	o	-	-	o	o	+	-	-	-	-	-	-	-	-
V_31H	+	-	-	-	o	o	+	-	-	-	-	-	-	-	-
V_33	o	o	o	-	o	o	o	-	-	+	o	o	-	-	-
V_32H	o	-	o	-	o	o	-	+	-	-	+	-	-	-	-

lglex

Table des classes des verbes $\longrightarrow$ possibilité d'extraire un **lexique syntaxique** semi-structuré des verbes simples et des noms prédicatifs à partir des tables [Constant & Tolone 2008] :

- format textuel ou XML
- nommé lexique **lglex**
- conversion depuis les tables Excel grâce à l'outil *LGExtract*

Extrait du lexique *lglex*

ID=V_35L_242
*lexical-info=[locs=(loc=[id="1",list=()],loc=[id="2",list=()]),cat="verb",verb=[***lemma="ruisseler"***],*
aux-list=(),prepositions=()]
args=(
const=[dist=(comp=[cat="NP",source="true",introd-prep=(),origine=(orig="Loc N1 =: de N1 source"),
introd-loc=(prep="de")]),pos="1"],
*const=[dist=(comp=[***cat="NP"***,introd-prep=(),origine=(orig="Loc N2 =: vers N2 destination",*
*orig="Loc N2 =: dans N2 destination"),***introd-loc=(prep="vers",prep="dans")***,destination="true"]),***pos="2"***],*
const=[pos="0",dist=(comp=[cat="NP",introd-prep=(),nothum="true",origine=(orig="N0 =: N-hum"),
introd-loc=()])])
*all-constructions=[absolute=(***construction="o::N0 V Loc N1 source Loc N2 destination"***,construction="o::N0 V",*
construction="o::N0 être V-ant",construction="true::N0 V Loc N1"),
relative=(construction="Ppv =: y",construction="Ppv =: en",construction="[extrap]")]
example=[example="L'eau ruisselle de la gouttière sur les passants"]

- Reste à faire : **interpréter** un certain nombre de colonnes

3. Conversion de *lglex* en un lexique pour le TAL

Le Lefff

- Le Lefff (Lexique des Formes Fléchies du Français) est un lexique morphologique et syntaxique pour le français [Sagot *et al.* 2006]
  $\longrightarrow$ utilisé par l'analyseur syntaxique FRMG :
  Ceci permet une comparaison entre $\text{FRMG}_{\text{Lefff}}$ et FRMG_{lglex}
- Le lexique **intensionnel**
  - associe à chaque entrée (emploi d'un lemme) un cadre de sous-catégorisation canonique
  - liste les redistributions possibles à partir de ce cadre
- Le processus de **compilation** du lexique intensionnel en lexique **extensionnel** construit différentes entrées pour chaque forme fléchie du lemme et chaque redistribution possible

Fonctions, réalisations et redistributions

- **Fonctions syntaxiques** (cf. Dicovalence) : Suj, Obj, Objà, Objde, Loc, Dloc, Att, Obl/Obl2
- **Réalisations** : directes (sn, sa, sinf, scompl, qcompl) ; clitiques (cln, cla, cld, y, en) ; prépositionnelles (prep+directe, par ex. par-sn, à-sinf, de-scompl)
- **Redistributions** : %actif, %passif, %passif_impersonnel...

Résultat de la conversion sur l'exemple précédent

ruisseler$^{35\mathrm{L}}_{242}$ *v-er:std*
100;Lemma;v;
*<***Suj***:cln|sn,***Dloc***:(de-sn|en),***Loc***:(vers-sn|dans-sn|y)>;*
cat=v;
%actif

- *L'eau ruissèle de la montagne vers la vallée*

Lexique obtenu : *lglex*$_{\text{Lefff}}$

- La conversion de *lglex* au format Lefff n'est **pas simple**
  $\longrightarrow$ effectuée pour les verbes simples [Tolone & Sagot 2009]
- Le lexique verbal obtenu, *lglex*$_{\text{Lefff}}$, contient 16 903 entrées pour 5 694 lemmes verbaux différents (2,96 entrées/lemme).
  - À titre de comparaison, le Lefff contient seulement 7 072 entrées verbales pour 6 818 lemmes verbaux distincts (1,04 entrées/lemme)

4. Intégration dans FRMG et évaluation

Intégration dans FRMG et évaluation

- Il suffit de remplacer les entrées verbales du $\mathrm{Le}\mathit{fff}$ par $\mathit{lglex}_{\mathrm{Le}\mathit{fff}}$
- Le résultat est une **variante de FRMG**, que nous noterons $\mathrm{FRMG}_{\mathit{lglex}}$, par opposition à la variante standard $\mathrm{FRMG}_{\mathrm{Le}\mathit{fff}}$
- Évaluation de $\mathrm{FRMG}_{\mathrm{Le}\mathit{fff}}$ et $\mathrm{FRMG}_{\mathit{lglex}}$ en *chunks* et en *relations* ($\sim$ dépendances entre mots pleins) sur la partie annotée manuellement du copus EASy
  $\longrightarrow$ selon les métriques de la la première campagne EASy d'évaluation des analyseurs syntaxiques [Paroubek *et al.* 2006]

Résultats

- FRMG$_{lglex}$ donne de meilleurs résultats que FRMG$_{\mathrm{Le}fff}$ pour **certains chunks** ou **certaines relations** en fonction des corpus, et inversement
  [Tolone & Sagot 2009]
- L'**ambiguïté lexicale** est plus élevée dans FRMG$_{lglex}$ puisque le nombre d'entrées est plus élevé. Cela induit :
  - une ambiguïté plus élevée dans l'analyseur
  - et donc d'autant plus de risque de se tromper au moment de la désambiguïsation

Conclusions et perspectives

Amélioration du lexique obtenu

- De nombreuses phrases reçoivent une **analyse complète** par FRMG$_{lglex}$ mais pas par FRMG$_{\mathrm{Le}fff}$, et inversement
  - $\rightarrow$ **coupler les 2 variantes de l'analyseur** pour garder un maximum d'analyses complètes, bien meilleures que les analyses partielles
- $\mathrm{Le}fff$ et $lglex_{\mathrm{Le}fff}$ étant **complémentaires**, il faut étudier les différences entre les différentes erreurs faites par chacune des variantes de l'analyseur
  - $\rightarrow$ utiliser des **techniques automatiques de fouilles d'erreurs** pour améliorer chaque ressource mutuellement (i.e., analyse statistique des résultats de l'analyse syntaxique [Sagot & de La Clergerie 2008])

Généralisation de la méthode

Optimiser l'utilisation des données lexicales du Lexique-Grammaire pour l'analyse syntaxique

- continuer à **améliorer les tables**
- finir la **table des classes** pour chaque catégorie
- **améliorer/corriger le processus de conversion**
- appliquer cette technique aux tables du Lexique-Grammaire pour les **autres catégories**
- généraliser cette méthode pour les **autres langues** pour lesquelles des tables du Lexique-Grammaire à large-couverture sont disponibles (i.e., Grec)

Annexe Résultats

Résultats comparatifs EASy de FRMG$_{\mathrm{Lefff}}$ et FRMG$_{lglex}$ (pourcentages de f-mesure) :

	Chunks		Relations	
Sous-corpus	FRMG$_{\mathrm{Lefff}}$	FRMG$_{lglex}$	FRMG$_{\mathrm{Lefff}}$	FRMG$_{lglex}$
`general_lemonde`	**86.8%**	82.8%	**59.8%**	56.9%
`general_senat`	82.7%	**83.1%**	**56.7%**	54.9%
`litteraire_2`	**84.7%**	81.5%	**59.2%**	56.3%
`medical_2`	85.4%	**89.2%**	**62.4%**	58.6%
`oral_delic_8`	**74.1%**	73.6%	47.2%	**48.5%**
`questions_amaryllis`	90.5%	**90.6%**	**65.6%**	63.2%
total	**84.4%**	*82.3%*	**59.9%**	*56.6%*

Temps d'analyse plus élevés avec FRMG$_{lglex}$ qu'avec FRMG$_{\mathrm{Lefff}}$: temps médian par phrase de 0,62 s contre 0,26 s

Annexe Précautions

Les résultats de FRMG$_{\mathit{lglex}}$ seront à interpréter en gardant à l'esprit plusieurs points

- FRMG utilise les entrées verbales converties à partir des tables, et non pas les entrées telles qu'elles sont dans les tables
  $\longrightarrow$ le processus de conversion est sûrement entâché d'erreurs
- le $\mathrm{Le}\mathit{fff}$ a été développé en parallèle aux campagnes EASy, contrairement aux tables
- $\mathit{lglex}_{\mathrm{Le}\mathit{fff}}$ ne contient pas toutes les entrées verbales nécessaires, et il a fallu en rajouter
  $\longrightarrow$ il se peut que certaines entrées manquent encore