Exploitation des tables du Lexique-Grammaire pour l'analyse syntaxique automatique

Elsa Tolone[1]

1. IGM, Université Paris-Est (France)

Séminaire de l'Équipe Informatique Linguistique
14 juin 2010

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 directement exploitables** telles quelles dans un analyseur syntaxique
  - les propriétés définitoires ne sont pas représentées
  - le format de représentation des informations lexicales n'est pas formalisé
  - il faut interfacer ces informations avec un analyseur syntaxique réel

Objectifs

Triple objectif :

- **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

Outils TAL retenus :

- analyseur syntaxique : FRMG [Thomasset & de La Clergerie 2005]
- formalisme lexical : Alexina, format du lexique Lefff [Sagot 2010] utilisé par FRMG

→ ceci permet une comparaison entre $\mathrm{FRMG}_{\mathrm{Lefff}}$ et FRMG_{LGLex}

Exemple de dépendances avec FRMG

Paul s'adresse à Max

→ entrée *se adresser_33_8 :v*

1. Les tables du Lexique-Grammaire
   1.1. Tables du Lexique-Grammaire
   1.2. Tables des classes
   1.3. Constructions de base
   1.4. *LGLex*
2. Le Lefff et le format Alexina
   2.1. Le Lefff
   2.2. Alexina
3. Conversion de *LGLex* en un lexique au format Alexina
   3.1. Identification des entrées à construire
   3.2. Construction des entrées
   3.3. Lexique obtenu : $LGLex_{\mathrm{Lefff}}$
4. Intégration dans l'analyseur syntaxique FRMG
5. Évaluation et discussion
   5.1. Protocole
   5.2. Résultats
   5.3. Discussion

1 Les tables du Lexique-Grammaire

Tables du Lexique-Grammaire

- Caractéristiques :
  - Développées **manuellement** depuis plus de 40 ans par le LADL, puis 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
- [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*)

Principe

- Chaque classe est codée dans une **table** :
  - lignes = entrées (mots)
  - colonnes = propriétés grammaticales
- **Propriétés** :
  - constructions
  - 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
  dont au moins une construction de base

Extrait de la table des verbes de la classe 33

N0 =: Nhum	N0 =: N-hum	N0 =: Nnr	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

Construction de base : N_0 V à N_1

Bilan

- Inventaire [Tolone 2009] :
  - 64 classes de **verbes simples**
    - 13862 entrées pour 5739 lemmes distincts
  - 32 classes d'**adverbes simples et figés** (adverbes en *-ment* et locutions adverbiales)
    - 10487 entrées pour 9273 lemmes distincts
  - 78 classes de **noms prédicatifs simples et figés** (noms avec argument(s) qui sont étudiés avec leur verbes support)
    ex : *Luc monte une attaque contre le fort*
    - 12696 entrées pour 8530 lemmes distincts
  - 69 classes d'**expressions figées**
    ex : *arriver à la cheville*
    - 39627 entrées pour 38605 lemmes distincts
- Lexique riche à large couverture → bientôt en version 3.2 contenant l'ensemble des tables distribuées actuelles (version 1 = format d'origine, version 2 = 1ères tables en ligne)

Entrées supplémentaires

- noms prédicatifs simples :
  - 24 classes codant également des **verbes simples**
    ex : *Max a donné des encouragements à Léa*
    → *Max encourage Léa*
    - 3334 entrées pour 2377 lemmes distincts
      → Éliminer les doublons pour ne pas créer plus d'ambiguïtés
  - 8 classes codant également des **adjectifs prédicatifs simples**
    ex : *Max a une certaine aptitude à supporter la douleur*
    → *Max est apte à supporter la douleur*
    - 1408 entrées pour 1287 lemmes distincts
- adverbes :
  - 2 classes codant également des **expressions figées verbales**
    ex : *partir comme une fusée*
    - 388 entrées pour 385 lemmes distincts
  - 1 classe codant également des **adjectifs prédicatifs composés**
    ex : *être doux comme un ange*
    - 304 entrées pour 300 lemmes distincts

Problèmes

- **Noms différents** pour une même propriété
  $\longrightarrow$ Harmonisation des intitulés de colonnes (tout OK)
  ex : *[extrap]* et *il V* N_0 *W*
- Propriétés **pas définies clairement**
  $\longrightarrow$ Documentation des propriétés (verbes, adverbes en cours)
- **Propriétés définitoires implicites** (littérature)
  $\longrightarrow$ Constante $+$ ou $-$ pour l'ensemble de la table (tout OK)
- Toutes les propriétés ne sont **pas codées** dans chaque table
  $\longrightarrow$ Codage $+$, $-$ ou *o* pour l'ensemble de la table (verbes en cours)
- Entrées pas **à jour** dans les tables
  $\longrightarrow$ Ajout des entrées manquantes (verbes OK)
- Entrées **non codées** ($\sim$)
  $\longrightarrow$ Codage d'entrée (verbes en cours)

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, bien avancé pour les verbes et les noms prédicatifs [Constant & Tolone 2010], mais aussi pour les expressions figées et les adverbes

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	-	+	-	-	+	-	-	-	-

Verbes simples : les complétives

- Modification de la construction de base

Table	Construction de base d'origine	Construction de base actuelle	Colonne codée dans la table
1	N0 U Prép V0-inf W	N0 V Prép V0-inf W	
3	N0 V N1 V1-inf W	N0 V N1 Loc N2 V1-inf W N0 V N1 V1-inf W N0 V N1 Loc N2	

- Éclatement en plusieurs classes

Table	Construction de base d'origine	Construction de base actuelle	Colonne codée dans la table
2	N0 V (E+N1) V0-inf	N0 V Loc N1 V0-inf W N0 V Loc N1 (*filer dehors acheter qqch*)	N0 V V0-inf W
2T		N0 V N1 V0-inf W (*franchir la rivière retrouver qqun*)	N0 V V0-inf W

Verbes simples : les complétives

- Changement de notation

Table	Construction de base d'origine	Construction de base actuelle	Colonne codée dans la table
4	Qu P V N1 (N0 = : Nnr codée +)	N0 V N1 N0 = : Nnr	
5 (12)	Qu P V Prép N1	N0 V Prép N1 N0 = : Qu P	
6 (7 8 9 10 13 11 14 15 18)	N0 V Qu P	N0 V Prép N1	N1 = : Qu Pind N1 = : Qu Psubj N1 = : si P ou si P N1 = : ce(ci+la) N1 = : Nhum N1 = : N-hum N1 = : le fait Qu P

Verbes simples : les complétives

- Suppression de tables

Table	Construction de base d'origine	Exemple
17	il V (E+Prép ce) Qu P (E+Prép N2)	*il s'agir pour ... de* *il apparaître à* → expressions figées (COQ) : Det0 C0 V Prép1 (Det1 C1+N1) Prép2 N2
19	Qu P V N1 Prép N2	*apporter* *remplir* → verbes de BGL (36DT,37M4) : N0 V N1 Prép N2

Noms prédicatifs : verbe support *avoir*

- Changement d'intitulés

Table	Construction de base	Intitulé de colonne d'origine	Intitulé de colonne actuel
AN09	N0 avoir Det N	N0 avoir un N N0 avoir un certain N N0 avoir des N	Det = : un Det = : un-certain Det = : des

- Modification de la construction de base

Table	Construction de base d'orgine	Construction de base actuelle	Colonne codée dans la table
AN07	N0 avoir Det N (*E+Modif) = N0 être de Det N (*E+Modif)	N0 avoir Det N Det = : un-Modif Det = : un-certain	N0 être de Det N Modif
AN08	N0 avoir Det N = il y avoir Det N Loc N0 avoir = comporter, comprendre	N0 avoir Det N Vsup = : comporter Vsup = : comprendre	il y avoir Det N Loc N0

Expressions figées

- Construction de base incluant la structure des expressions figées (N : argument libre, C : argument figé)

Table	Construction de base	Exemple
EPCPN	N0 être Prép1 Det1 C1 Prép2 N2	*être à la frontière entre*

Adverbes

- Construction de base des adverbes de manière

Table	Construction de base	Exemple
ADVM...	N0 V Adv W	*Ce livre est en vente exclusivement sur ce site* **Exclusivement, ce livre est en vente sur ce site*
ADVMP ADVMS ADVMTF	Adv, N0 V W	*Ce livre est en vente régulièrement sur ce site* *Régulièrement, ce livre est en vente sur ce site* **Régulièrement, ce livre n'est pas en vente sur ce site*
ADVMP	Adv, N0 ne V pas W	*Ce concert est musicalement une réussite* *Musicalement, ce concert est une réussite* *Musicalement, ce concert n'est pas une réussite*

Adverbes

- Construction de base des adverbes de phrase

Table	Construction de base	Exemple
ADVP...	Adv, N0 V W Adv, N0 ne V pas W	*Cinquièmement, Sunrider n'utilise aucun* sous-produit animal **Sunrider n'utilise cinquièmement aucun* sous-produit animal
ADVPC	P1 Adv P2	

- Construction de base des adverbes figés
  + Structure des adverbes figés

Table	Construction de base	Exemple
P...	N0 V Adv W Adv, N0 V W	
	Adv, N0 ne V pas W	
PCPC	Prép1 Det1 C1 Prép2 Det2 C2	*changer de le jour à le lendemain*

LGLex

Table des classes → possibilité d'extraire un **lexique syntaxique** semi-structuré pour chaque catégorie à partir des tables [Constant & Tolone 2010] :

- format textuel ou XML
- nommé lexique *LGLex*
- conversion depuis les tables Excel grâce à l'outil *LGExtract*
- Version 3.2 contenant l'ensemble des tables distribuées bientôt disponible sur `http://infolingu.univ-mlv.fr/`

LGLex est le point de départ du processus de conversion vers le format Alexina

Format du lexique *LGLex*

- **ID=catégorie_numTable_numEntrée**
- **lexical-info**=[...] → lemme et ses informations lexicales (auxiliaires, verbes supports, déterminants, prépositions)
- **args**=(...) → arguments et leurs distributions accompagnées d'autres informations (traits sémantiques, mode de la complétive, argument contrôlé par l'infinitive, prépositions)
- **all-constructions**=[absolute=(...), relative=(...)] → liste des constructions acceptées (soit nommées de façon complète avec tous les éléments dans l'ordre, soit transformations à partir de la construction de base, mais aussi réductions de GN, structures des adverbes figés,. . .)
- **example**=[...] → exemple illustrant l'entrée

Extrait du lexique *LGLex* (verbe)

ID=V_35L_242
*lexical-info=[cat="verb",verb=[***lemma="ruisseler"***],*
locs=(loc=[id="1",list=()],loc=[id="2",list=()]),aux-list=(),prepositions=()]
args=(
const=[pos="0",dist=(comp=[cat="NP",nothum="true",introd-loc=(),introd-prep=(),
origine=(orig="N0 =: N-hum")])],
*const=[***pos="1"***,dist=(comp=[***cat="NP"***,source="true",***introd-loc=(prep="de")***
introd-prep=(),origine=(orig="Loc N1 =: de N1 source")])],
*const=[***pos="2"***,dist=(comp=[***cat="NP"***,destination="true",***introd-loc=(prep="vers",prep="dans")***,*
introd-prep=(),origine=(orig="Loc N2 =: vers N2 destination",
orig="Loc N2 =: dans N2 destination")])])
*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

2. Le Lefff et le format Alexina

Le Lefff

- Le Lefff (Lexique des Formes Fléchies du Français) est un lexique morphologique et syntaxique pour le français [Sagot 2010]
  - à large couverture (536,375 entrées correspondant à 110,477 lemmes distincts couvrant toutes les catégories)
  - librement distribué (licence LGPL-LR)
- Il repose sur l'architecture **Alexina** d'acquisition et de modélisation de lexiques morphologiques et syntaxiques

Alexina

Architecture à deux niveaux

- 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

Alexina sur un exemple

- Exemple d'entrée intentionnelle :

clarifier$_1$ *v-er:std*
Lemma;v;
<**Suj**:*cln*|*scompl*|*sinf*|*sn*,**Obj**:(*cla*|*scompl*|*sn*)>;
%ppp_employé_comme_adj,*%actif*,*%se_moyen_impersonnel*,
%passif_impersonnel,*%passif*

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, p.ex. par-sn, à-sinf, de-scompl)
- **Redistributions** : %actif, %passif, %se_neutre, %actif_impersonnel,...

3. Conversion de *LGLex* en un lexique au format Alexina

Différents types de constructions

Chaque entrée de *LGLex* est associée à des constructions dont on peut distinguer plusieurs types :

1. la construction **de base**, définitoire de la classe de l'entrée
2. les constructions **de base étendues**, obtenues par adjonction d'arguments à la construction de base
   - en pratique, ces constructions sont toutes des intermédiaires entre la construction de base et une construction dite **de base maximalement étendue**
3. les **variantes** de la construction de base, obtenues par effacement d'un ou de plusieurs arguments ou par changement de type de réalisation
4. les constructions qui sont des **redistributions**
5. les constructions dont il semble qu'elles auraient dû conduire à des entrées distinctes, dites **entrées secondaires**

Identification des entrées à construire

- Nous avons développé une méthode permettant d'*aligner* deux constructions [Tolone & Sagot 2009]
- $\rightarrow$ identification de la construction de base maximalement étendue et de ses variantes
  - ces constructions sont rassemblées en une seule entrée dite **entrée canonique**
- Autres constructions :
  - certaines correspondent à des **redistributions standard** ([passif par], [extrap],...) $\rightarrow$ ajout à l'entrée canonique de cette redistribution
  - les autres induisent la création d'**entrées distinctes**

Construction d'une entrée

Pour chaque entrée à construire :

- on extrait la **sous-catégorisation du cadre maximal**
  - les **fonctions syntaxiques** sont calculées à l'aide d'heuristiques utilisant la position, la nature et certaines propriétés des arguments
  - les **réalisations** sont obtenues soit directement, soit par l'intermédiaire des champs *prep* et *loc* lorsque la construction mentionne un argument en *Prép X*, soit grâce à des propriétés spécifiques (par exemple, *à* $N_1 = Ppv = : le$)
- on prend en compte les différentes **variantes** du cadre maximal en rendant **facultatifs** les arguments appropriés
- on ajoute des informations complémentaires lorsqu'on peut les extraire de propriétés
  - informations de contrôle, de mode pour les complétives,. . .

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 du haut de la montagne vers le bas de la vallée*
  - *L'eau ruissèle du haut de la montagne*
  - *L'eau ruissèle vers le bas de la vallée*
  - *L'eau ruissèle*

Autre extrait du lexique *LGLex* (nom)

ID=N_aa_18
*lexical-info=[cat="noun",noun=[notperm=[***complete="considération"***],noun1="considération"]],*
Vsup=[cat="verb",list=(value="ressentir",value="éprouver",value="avoir")],
Vsup2=[cat="verb",list=(value="avoir",value="être l'objet de",value="faire l'objet de")],
detN=[list-det-modif=(det-modif=[det="du+de l'+de la",modif="false"],
*det-modif=[det="un+une",modif="true"]),prepositions=(***preposition=[id="1",list=(prep="pour")]***)]*
args=(
const=[pos="0",dist=(comp=[cat="NP",hum="true"])],
*const=[***pos="1"***,dist=(comp=[cat="NP",nothum="true"],comp=[cat="NP",hum="true"])])*
*all-constructions=[absolute=(***construction="true::N1 avoir Det N de N0"***,*
construction="true::N0 avoir Det N Prép N1"),
reductionsGN=(construction="::le N de N0 Prép N1"),relative=(),verbales=()]
example=[example="Max a de la considération pour (ce geste+Luc)"]

Résultat de la conversion

considération $^{\mathrm{aa}}_{18}$ *nc-2f*
100;Lemma;cf;
<**Suj**:*cln*|*sn*,**Obl**:*pour-sn*>;
cat=nc;
lightverb=avoir|*ressentir*|*éprouver;*
%default

considération $^{\mathrm{aa}}_{18}$ *nc-2f*
100;Lemma;cf;
<**Suj**:*cln*|*sn*,**Objde**:*de-sn*>;
cat=nc;
lightverb=avoir;
%default

Version 3.2 contenant les verbes et les noms des tables distribuées bientôt disponible sur `http://infolingu.univ-mlv.fr/`

Autre extrait du lexique *LGLex* (expression figée, adverbe)

ID=C_a1_53
*lexical-info=[cat="expr",notperm=[***complete="<E> avoir une bonne bille <E> <E> <E>"***],*
expr=[pos="1",carg1=[verb=[cat="verb",values=(value="avoir")],det1=[cat="det",value="une bonne"],
c1=[cat="noun",value="bille"],detc=[cat="detc",value="<E>"],
prepc=[cat="prep",value="<E>"],cc=[cat="noun",value="<E>"]]]]
args=(const=[pos="0",dist=(comp=[cat="NP",hum="true",introd-loc=(),introd-prep=(),origin=()])])
*all-constructions=[absolute=(***construction="true::N0 avoir Det1 C1 Prépc Detc Cc"***),relative=()]*
example="§a§une bonne bille"

ID=P_advmf_1
*lexical-info=[cat="adv",adv=[notperm=[***complete="approximativement"***],*
carg1=[pos="1",adv=[cat="adv",value="approximativement"]]]]
args=()
*all-constructions=[absolute=(***construction="true::N0 V Adv W"***,***construction="o::Adv, N0 V W"***),*
structureAdv=(),relative=(construction="très Adv")

- Reste à faire : convertir au format Lefff

Lexique obtenu : $LGLex_{\mathrm{Lefff}}$

- **verbes** : $LGLex_{\mathrm{Lefff}}$, contient 16 955 entrées pour 5 723 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)

  Au niveau extensionnel, le Lefff contient 361 268 entrées, alors que le lexique extrait de *LGLex* en contient 862 546
- **noms prédicatifs** : $LGLex_{\mathrm{Lefff}}$ contient 17 517 entrées pour 8 363 lemmes différents (2,09 entrées/lemme)
  - Le Lefff ne contient que 218 entrées de noms prédicatifs (1 entrée/lemme)

  Au niveau extensionnel, *LGLex* contient 32 389 entrées

4. Intégration dans l'analyseur syntaxique FRMG

Intégration dans l'analyseur syntaxique FRMG

- L'intégration de $LGLex_{\mathrm{Lefff}}$ dans l'analyseur FRMG est rapide
- Le *lexeur* de l'analyseur fait normalement appel à une base de données lexicales construite à partir du Lefff
- Nous allons donc construire un autre lexique, qui repose sur $LGLex_{\mathrm{Lefff}}$, que nous demanderons au lexeur d'utiliser à la place du Lefff

Pour le moment, intégration uniquement des verbes, car les résultats ne sont pas encore probants avec les noms prédicatifs

Intégration dans l'analyseur syntaxique FRMG

Il faut donc :

- remplacer les entrées verbales du Lefff par $LGLex_{\text{Lefff}}$
- conserver les autres entrées du Lefff
- compléter le résultat par diverses entrées verbales venant du Lefff, qui ne font pas partie du lexique *LGLex*
  - entrées pour les auxiliaires et semi-auxiliaires
  - certains verbes à montée
  - les verbes impersonnels
  - les entrées pour les têtes syntaxiques des constructions à verbes support
- construire la base de données lexicales correspondantes
- spécifier à FRMG d'utiliser cette dernière

Le résultat est une variante de l'analyseur FRMG, que nous noterons FRMG_{LGLex}, par opposition à la variante standard notée $\text{FRMG}_{\text{Lefff}}$

5.1. Protocole
5.2. Résultats
5.3. Discussion

5. Évaluation et discussion

Protocole utilisé

- Nous avons évalué $\text{FRMG}_{\text{Lefff}}$ et FRMG_{LGLex} en analysant la partie annotée manuellement du corpus EASy [Paroubek *et al.* 2005]
  - 4 306 phrases de styles variés (journalistique, médical, oral, questions, littéraire,. . .)
- métriques utilisées : celles de la première campagne EASy d'évaluation des analyseurs syntaxiques, qui a eu lieu fin 2005 [Paroubek *et al.* 2006]
  - évaluation en *chunks* et en *relations* ($\sim$ dépendances entre mots pleins)

Précautions

Les résultats de FRMG_{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
- EASy ne permet pas d'évaluer toutes les informations contenues dans les tables (informations sémantiques par ex.)
- le Lefff a été développé en parallèle aux campagnes EASy, contrairement aux tables
- $LGLex_{\text{Lefff}}$ ne contient pas toutes les entrées verbales nécessaires, et il a fallu en rajouter
  $\longrightarrow$ il se peut que certaines entrées manquent encore

Résultats

Résultats comparatifs EASy de $\text{FRMG}_{\text{Lefff}}$ et FRMG_{LGLex} (pourcentages de f-mesure) :

	Chunks		Relations	
Sous-corpus	$\text{FRMG}_{\text{Lefff}}$	FRMG_{LGLex}	$\text{FRMG}_{\text{Lefff}}$	FRMG_{LGLex}
`general_lemonde`	**88.22%**	84.60%	**62.73%**	59.01%
`litteraire_2`	**88.91%**	88.46%	**65.28%**	62.43%
`mail_9`	**82.60%**	81.90%	**58.55%**	56.00%
`medical_3`	85.04%	**85.89%**	64.79%	**65.26%**
`oral_delic_4`	78.80%	**81.79%**	**51.67%**	51.14%
`questions_amaryllis`	**91.30%**	90.73%	**66.56%**	64.77%
total	**87.05%**	*85.53%*	**63.10%**	*60.25%*

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

- ceci provient du nombre d'entrées par lemme 3 plus élevé dans *LGLex* que dans le Lefff

- FRMG_{LGLex} donne de meilleurs résultats que $\text{FRMG}_{\text{Lefff}}$ pour certaines relations
  - relations « classiques » MOD_A (modifieur-adjectif) et MOD_R (modifieur-adverbe)
  - relations « difficiles » MOD_P (modifieur-préposition) et APP (apposition)
- la relation ATB_SO (attribut du sujet ou de l'objet) est celle pour laquelle la différence en rappel est la plus importante (34,0% contre 58,4%) ;
  - les tables du Lexique-Grammaire encodent très peu d'informations sur les phénomènes d'attribution, mais cela peut aussi être dû aux erreurs de la référence EASy

Exemple : annotation des constituants

Exemple : annotation des relations

Relations

1. Sujet - Verbe

sujet	verbe
G2 F5 67%	G5 F9

2. Auxiliaire - Verbe

auxiliaire	verbe
G5 F9	G6 F10

3. COD - Verbe

COD	verbe
G7	G6

4. Complément - Verbe

complément	verbe
G1 F2 67%	G6 F10

7. Attribut - Sujet / Objet

attribut	verbe	sujet / objet
F11	F10	sujet

8. Modifieur - Nom

modifieur	nom	à propager
G3 F7 67%	G2 F5 67%	
F4	F5	
G4 F8	G3 F7 67%	
G8 F13 67%	G7 F11	
G9 F14	G8 F13 67%	
G11 F19 67%	G10 F17 50%	
G10	G7	
F17	F13	

- l'**ambiguïté lexicale** plus élevée dans FRMG_{LGLex} induit
  - une ambiguïté plus élevée dans l'analyseur
  - et donc d'autant plus de risque de se tromper au moment de la désambiguïsation
- exemple :
  - *[...] on estime que cette décision [ferait] dérailler le processus de paix*
  - FRMG utilise l'heuristique habituelle « on préfère les arguments aux modifieurs »
  - FRMG_{LGLex} fait de *de paix* un argument de *estimer* (*estimer qqch de qqn*)
  - $\text{FRMG}_{\text{Lefff}}$ ne se trompe pas car dans le Lefff, *estimer* n'a pas d'Objde

Conclusions et perspectives

À court terme

- De nombreuses phrases reçoivent une analyse complète par FRMG_{LGLex} mais pas par $\text{FRMG}_{\text{Lefff}}$, et inversement
  - par exemple, sur le sous-corpus `general_lemonde`, 177 phrases sont entièrement reconnues par les deux analyseurs, 85 seulement par $\text{FRMG}_{\text{Lefff}}$, 76 seulement par FRMG_{LGLex}, et 111 par aucun des deux
- Les analyses complètes sont souvent bien meilleures que les analyses partielles
- Possibilité d'utiliser les deux analyseurs en même temps, et de ne conserver le résultat que de celui qui a donné une analyse complète

À long terme

Exploiter la complémentarité des deux ressources pour y détecter des erreurs et les améliorer mutuellement

- étudier les différences entre les erreurs faites par chacun des deux analyseurs
- utiliser des techniques automatiques de fouille d'erreurs pour extraire automatiquement des entrées probablement erronées [Sagot et de La Clergerie 2008]
  - pour $LGLex_{\mathrm{Lefff}}$, de nombreuses erreurs que l'on trouverait viendraient probablement du processus de conversion
  - certaines erreurs détectées pourraient provenir d'erreurs dans les tables, que l'on pourrait ainsi corriger

Tester le lexique sur un analyseur syntaxique probabiliste en cours de développement à l'IGM (Anthony Sigogne)

Généralisation de la méthode

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

- coder les **entrées manquantes ou non codées** dans les tables
  - coder en priorité les entrées plus courantes
  - coder en priorité les constructions les plus longues
- coder 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 en cours)

Références

- [Constant & Tolone 2010] Matthieu Constant et Elsa Tolone. A generic tool to generate a lexicon for NLP from Lexicon- Grammar tables. Lingue d'Europa e del Mediterraneo, Grammatica comparata, vol.1, pages 79-93. Aracne. 2010.
- [Gross 1975] Maurice Gross. Méthodes en syntaxe : Régime des constructions complétives. Hermann. Paris, France. 1975.
- [Paroubek et al. 2005] Patrick Paroubek, Louis-Gabriel Pouillot, Isabelle Robba et Anne Vilnat. Premier bilan de la participation du LORIA à la campagne d'évaluation EASy. Actes de l'atelier EASy de TALN'05. Dourdan, France. 2005.
- [Paroubek et al. 2006] Patrick Paroubek, Isabelle Robba, Anne Vilnat et Christelle Ayache. Data, Annotations and Measures in EASy : the Evaluation Campaign for Parsers of French. Actes de LREC'06. Gênes, Italie. 2006.

Références (suite)

- [Sagot 2010] Benoît Sagot. The Lefff, a freely available and large-coverage morphological and syntactic lexicon for French. Actes de LREC'10, 8 pages. La Valette, Malte. 2010.
- [Sagot et de La Clergerie 2008] Benoît Sagot et Éric de La Clergerie. Fouille d'erreurs sur les sorties d'analyseurs syntaxiques. Traitement Automatique des Langues (T.A.L.), vol.49, num.1, p.41-60. Hermès. Paris, France. 2008.
- [Thomasset & de La Clergerie 2005] François Thomasset and Éric de La Clergerie. Comment obtenir plus des méta- grammaires. Actes de TALN'05. Dourdan, France. 2005.
- [Tolone 2009] Elsa Tolone. Les tables du Lexique-Grammaire au format TAL. Actes de MajecSTIC 2009. Avignon, France. 2009.
- [Tolone & Sagot 2009] Elsa Tolone et Benoît Sagot. Using Lexicon-Grammar tables for French verbs in a large-coverage parser. Actes de LTC'09, pages 200-204. Poznań, Pologne. 2009.