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

Elsa Tolone[1]

1. LIGM, Université Paris-Est (France)

Séminaire de TALEP, LIF de Marseille, France
24 mars 2011

Contexte

- Les **tables du Lexique-Grammaire** du français = ressource lexicale très riche
  $\rightarrow$ informations **syntaxiques** et syntaxico-sémantiques
- Mais les tables ne sont **pas directement exploitables** telles quelles dans un analyseur syntaxique
  - informations importantes non représentées car implicites pour une table donnée
  - format de représentation des informations lexicales non formalisé
  - il faut interfacer ces informations avec un analyseur syntaxique symbolique réel

Objectifs

Triple objectif :

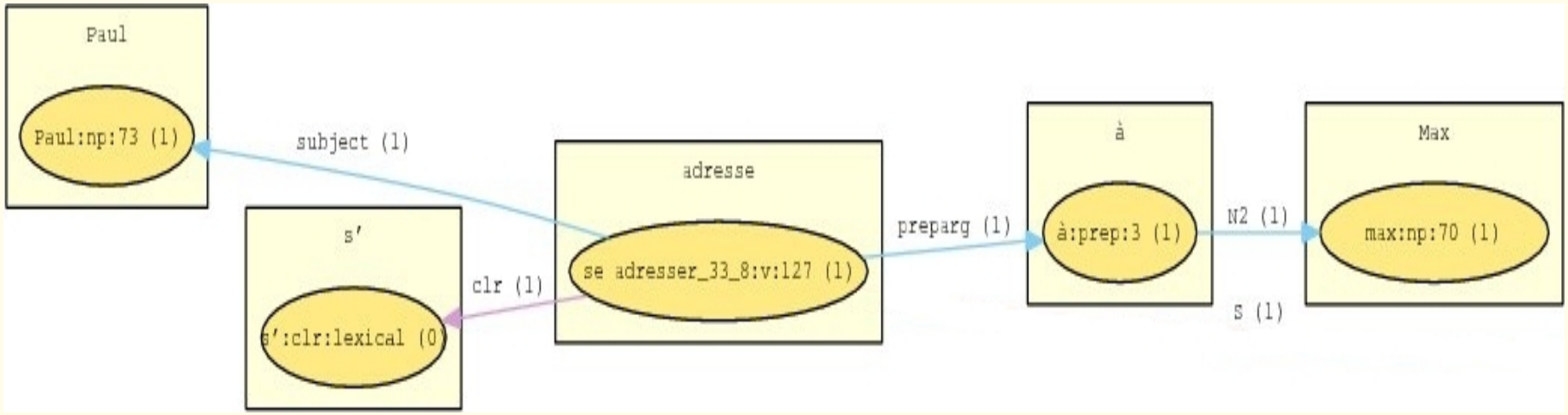
- **convertir** les tables du Lexique-Grammaire en un format TAL
- **coupler** le lexique syntaxique obtenu, nommé $LGLex_{\mathrm{Lefff}}$, avec un analyseur syntaxique
- **évaluer** l'analyseur obtenu

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 *s'adresser V_33_8*

1. Les tables du Lexique-Grammaire

2. Le lexique *LGLex*

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

4. Intégration dans l'analyseur syntaxique FRMG

5. Évaluation et discussion

1 Les tables du Lexique-Grammaire

Tables du Lexique-Grammaire

Développées **manuellement** depuis plus de 40 ans par le LADL [Gross 1975], puis l'Equipe d'Informatique Linguistique du LIGM (Université Paris-Est)

- Étude de la syntaxe d'une phrase élémentaire (ou **cadre de sous-catégorisation**)
  ex : N0 V à N1
- Catégories concernées : verbes, adverbes, noms et adjectifs prédicatifs et expressions figées pour le français
  $\rightarrow$ certaines **propriétés en commun** = **classes**
- 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, etc.)
    ex : N0 = : Nhum → nom humain
- Chaque propriété **testée pour chaque entrée**
  → codage binaire :
  - + : propriété acceptée
  - − : propriété non acceptée
- **Propriétés définitoires** pour chaque classe
  dont 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 : N0 V à N1

Bilan

- Inventaire [Tolone 2009] :
  - 67 classes de **verbes simples**
    - 13 867 entrées pour 5 738 lemmes distincts
  - 78 classes de **noms prédicatifs simples et composés** (noms avec argument(s) qui sont étudiés avec leur verbes support)
    - 12 696 entrées pour 8 531 lemmes distincts
      ex : *Luc monte une attaque contre le fort*
  - 69 classes d'**expressions figées, principalement verbales et adjectivales**
    - 39 628 entrées pour 38 658 lemmes distincts
      ex : *arriver à la cheville de qq'un*, *être à la retraite*
  - 32 classes d'**adverbes simples et (semi-)figés** (adverbes en *-ment* et locutions adverbiales)
    - 10 488 entrées pour 9 326 lemmes distincts
      ex : *[changer] du jour au lendemain*

Entrées supplémentaires

- noms prédicatifs simples :
  - 24 classes codant également des **verbes simples**
    - 3 334 entrées pour 2 377 lemmes distincts
      ex : *Max a donné des encouragements à Léa*
      → *Max encourage Léa*
  - 8 classes codant également des **adjectifs prédicatifs simples**
    - 1 408 entrées pour 1 287 lemmes distincts
      ex : *Max a une certaine aptitude à supporter la douleur*
      → *Max est apte à supporter la douleur*
- adverbes :
  - 2 classes codant également des **expressions figées verbales**
    - 388 entrées pour 385 lemmes distincts
      ex : *partir comme une fusée*
  - 1 classe codant également des **adjectifs prédicatifs composés**
    - 304 entrées pour 300 lemmes distincts
      ex : *être doux comme un ange*

Versions

Lexique riche à large couverture → la version 3.2 contient les tables distribuées actuelles pour chaque catégorie

- version 1 = format d'origine
- version 2 = 1ères tables en ligne (60% seulement)
- version 3 = depuis ma thèse, dans les différents formats de lexique
  - bientôt, la version 3.3 avec 100% des tables et pour les verbes :
    - documentation de toutes les propriétés
    - formules définitoires de chaque table
    - arbre de classification → déterminer à quelle classe appartient une entrée verbale donnée

Avantage des tables : lisibilité pour les linguistes

Problèmes

- **Noms différents** pour une même propriété
  → Harmonisation des intitulés de colonnes
  ex : [extrap] et il V N0 W
- Propriétés **pas définies clairement**
  → Documentation des propriétés
- **Propriétés définitoires implicites** (littérature)
  → Constante + ou − pour l'ensemble de la table
- Entrées pas **à jour** dans les tables
  → Ajout des entrées manquantes
- Entrées **non codées** (∼)
  → Codage d'entrée

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 harmonisation)
- 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

Table des classes créée pour chaque catégorie [Tolone et al. 2010]

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 de BGL

Cas des complétives et des infinitives [Tolone 2010]

- Simplification des constructions de base
- Exploitation informatique
- Spécification des distributions

Table	Construction de base d'origine	Construction de base actuelle	Autres propriétés définitoires
32CV	N0 V N1 (E+en V-n)	N0 V N1 en N2	N2 être V-n N0 V N1 N2 apparition

- *Max a roulé les papiers*
- *Max a roulé les papiers en un mince rouleau*

[Boons et al. 1976a]

Verbes de BGL (2)

- Éclatement en plusieurs classes [Tolone et al. 2010]

Table	Construction de base d'origine	Construction de base actuelle
35L	N0 V Loc N1	N0 V Loc N1 source Loc N2 destination

- *Paul a bondi du tabouret sur la table*
- *Paul a bondi du tabouret*
- *Paul a bondi sur la table*

Table	Construction de base actuelle
35LS	N0 V Loc N1 source

- *Le train a déraillé de la voie*

Table	Construction de base actuelle
35LD	N0 V Loc N1 destination

- *Le bateau s'enfonce dans les flots*

[Boons et al. 1976a]

Expressions figées (M. Gross)

- Reconstitution des constructions de base à partir des colonnes lexicales [Tolone et al. 2010]
- Modification des intitulés de colonnes pour être harmonisés avec la construction de base
- Structure des expressions figées incluse dans la construction de base (N : argument libre, C : argument figé)

Table	Construction de base
EPCPN	N0 être Prép1 Det1 C1 Prép2 N2

<ENT>être	<ENT>Prép1	<ENT>Det1	<ENT>C1	<ENT>Prép2
être	*en*	*<E>*	*accord*	*avec*
être	*de*	*la même*	*force*	*que*
être	*à*	*la*	*frontière*	*entre*

Table: Extrait de la table EPCPN des expressions figées (colonnes lexicales)

Adverbes en *-ment*

- Clarification des constructions de base [Tolone et al. 2010]
  - colonnes entièrement codées + supprimées

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*

[Molinier & Levrier 2000]

Adverbes (semi-)figés

- Reconstitution de la structure de l'adverbe à partir des colonnes lexicales [Tolone et al. 2010]
- Modification des intitulés de colonnes pour être harmonisés avec la construction de base

Table	Structure de l'adverbe	Exemple
PCPC	Prép1 Det1 C1 Prép2 Det2 C2	*[changer] du jour au lendemain*

- Définition des constructions de base (approximation)

Table	Construction de base
P...	N0 V Adv W Adv, N0 V W Adv, N0 ne V pas W

[Gross 1986]

2. Le lexique *LGLex*

LGLex

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

- nommé lexique *LGLex*
- conversion depuis les tables Excel grâce à l'outil *LGExtract*
- format d'échange, reposant sur les mêmes concepts linguistiques que ceux qui sont manipulés dans les tables
- format textuel ou XML
- version 3.2 contenant *LGLex* des tables distribuées pour chaque catégorie 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, etc.)
- **example**=[...] → exemple illustrant l'entrée

Extrait du lexique *LGLex* (verbe)

ID=V_36SL_28
*lexical-info=[cat="verb",verb=[***lemma="clouer"***],prepositions=(***preposition=[id="2",list=(prep="avec")]***)*
locatifs=(locatif=[id="2",list=()],aux-list=())]
args=(
*const=[***pos="0"***,dist=(comp=[***cat="NP",hum="true"***,introd-prep=(),introd-loc=(),*
origin=(orig="N0 = : Nhum")],
*const=[***pos="1"***,dist=(comp=[***cat="NP",nothum="true"***,introd-prep=(),introd-loc=(),*
origin=(orig="N1 = : N-hum")])])
*const=[***pos="2"***,dist=(comp=[***cat="NP",destination="true"***,introd-prep=(),***introd-loc=(prep="à",prep="sur")***,*
origin=(orig="Loc N2 = : à N2 destination",orig="Loc N2 = : sur N2 destination")])],
*all-constructions=[absolute=(construction="true : :***N0 V N1 Loc N2***",construction="o : :***N0 V N1***",*
construction="true : :N0 V N1 Prép N2", construction="true : :N0 V N1 et N2"
construction="o : :N0 V N1 de N3 attache",construction="o : :N0 V N1hum Loc N2abs",
construction="o : :N3 attache V N1",construction="o : :N0 V N1 + 2"),
*relative=(construction="***[passif par]***")]*
example=[example="Max a cloué cette planche(avec+contre+sur)celle-là"]

- entrée *clouer* V_36SL_28
- Reste à faire : **interpréter** les constructions

3. Conversion de *LGLex* en un lexique au 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 de modélisation et d'acquisition 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*
  100;Lemma;v;
  *<***Suj***:cln|scompl|sinf|sn,***Obj***:(cla|scompl|sn)>;*
  %ppp_employé_comme_adj,%actif,%se_moyen_impersonnel,
  %passif_impersonnel,%passif
- **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, etc.

Origine du Lefff

- acquisition automatique (avec validation manuelle) :
  - techniques statistiques appliquées sur des corpus bruts
  - informations syntaxiques atomiques
- correction/ajout manuel ou guidé par des techniques automatiques :
  - statistiques simples sur des corpus étiquetés
  - fouille d'erreurs dans les sorties d'analyseurs syntaxiques
- étude linguistique de certains phénomènes :
  - constructions impersonnelles, constructions pronominales, adverbes en *-ment*, plusieurs classes d'expressions figées verbales, verbes en *-iser* et *-fier*

$\rightarrow$ conversion de leur représentation dans d'autres ressources + fusion automatique (avec validation manuelle) avec le Lefff :

- tables du Lexique-Grammaire, Dicovalence, LVF + certains noms et adjectifs du lexique morphologique Multext

Le processus de conversion

- La conversion des tables du Lexique-Grammaire vers le formalisme Alexina n'est **pas directe** [Tolone & Sagot 2011]
  - Cela requiert une **definition formelle** ou une **interprétation dynamique** de tous les noms de propriétés
  - Directement ou indirectement, ces propriétés peuvent :
    - specifier un cadre de sous-catégorisation complet
    - spécifier une information partielle à propos d'un cadre de sous-catégorisation (le fait qu'un argument n'est pas obligatoire, une réalisation possible pour un argument, etc.)
    - correspondre à une redistribution
    - mener à une construction d'une entrée additionnelle
  - Des informations additionnelles importantes doivent être complétées grâce à des heuristiques ou à partir d'autres ressources lexicales
    - le nom de chaque fonction syntaxique, les phénomènes d'attribution, les informations morphologiques, etc.

L'exemple précédent du lexique *LGLex* (verbe)

ID=V_36SL_28
*lexical-info=[cat="verb",verb=[***lemma="clouer"***],prepositions=(***preposition=[id="2",list=(prep="avec")]***)*
locatifs=(locatif=[id="2",list=()]),aux-list=()]
args=(
*const=[***pos="0"***,dist=(comp=[***cat="NP",hum="true"***,introd-prep=(),introd-loc=(),*
origin=(orig="N0 = : Nhum")],
comp=[cat="NP",introd-prep=(),introd-loc=(),nothum="true",
origin=(orig="N0 = : N-hum")])],
*const=[***pos="1"***,dist=(comp=[***cat="NP",nothum="true"***,introd-prep=(),introd-loc=(),*
origin=(orig="N1 = : N-hum")])])
*const=[***pos="2"***,dist=(comp=[***cat="NP",destination="true"***,introd-prep=(),***introd-loc=(prep="à",prep="sur")***,*
origin=(orig="Loc N2 = : à N2 destination",orig="Loc N2 = : sur N2 destination")])],
*all-constructions=[absolute=(construction="true : :***N0 V N1 Loc N2***",construction="o : :***N0 V N1***",*
construction="true : :N0 V N1 Prép N2", construction="true : :N0 V N1 et N2"
construction="o : :N0 V N1 de N3 attache",construction="o : :N0 V N1hum Loc N2abs",
construction="o : :N3 attache V N1",construction="o : :N0 V N1 + 2"),
*relative=(construction='***[passif par]***")]*
example=[example="Max a cloué cette planche(avec+contre+sur)celle-là"]

- entrée *clouer* V_36SL_28

Résultat de la conversion sur l'exemple précédent

clouer V_36SL_28 *v-er :std*
100 ;Lemma ;v ;
*<***Suj** *:cln|sn,***Obj** *:sn,***Loc** *:(à-sn|sur-sn|avec-sn)> ;*
*cat=v ;***@SujNhum** *;***@ObjN-hum** *;*
%actif,%passif,%ppp_employé_comme_adj

- *Max a cloué ces planches au mur*
  - *Max a cloué ces planches*

Résultat de la conversion sur l'exemple précédent (2)

clouer V_36SL_28 *v-er :std*
100 ;Lemma ;v ;
*<**Suj** :cln|sn,**Obj** :sn,**Loc** :(à-sn|sur-sn|avec-sn)> ;*
cat=v ;@SujNhum ;@ObjN-hum ;
%actif,%passif,%ppp_employé_comme_adj

Entrée du lexique **intensionnel** :

- identifiant de l'entrée : catégorie_numTable_numEntrée
- classe morphologique, qui définit le modèle qui construit les formes fléchies en reposant sur les classes flexionnelles du *Lefff*
- poids
- catégorie
- cadre de sous-catégorisation initial
- informations supplémentaires représentées par des macros
- liste des redistributions possibles

Résultat de la conversion sur l'exemple précédent (3)

clouer V_36SL_28 *v-er :std*
100 ;Lemma ;v ;
*<***Suj** *:cln|sn,***Obj** *:sn,***Loc** *:(à-sn|sur-sn|avec-sn)> ;*
cat=v ;@SujNhum ;@ObjN-hum ;
*%actif,***%passif***,%ppp_employé_comme_adj*

Une entrée du lexique **extensionnel** :

clouées *100 v*
pred=clouer V_36SL_28
*<***Suj** *:sn,***Loc** *:(à-sn|sur-sn|avec-sn),***Obl** *:(par-sn)> ;*
@passive,**@pers**,*cat=v*,**@Kfp**

- *Ces planches ont été clouées au mur par Max*

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::***N0 avoir Det N Prép N1***"),*
*construction="true::***N1 avoir Det N de N0***",*
reductionsGN=(construction="::le N de N0 Prép N1"),relative=(),verbales=()]
example=[example="Max a de la considération pour (ce geste+Luc)"]

- *Max a de la* **considération** *pour (ce geste+Luc)*
- *(Ce geste+Luc) a la considération de Max*

Résultat de la conversion

considération N_aa_18 *nc-2f*
100;Lemma;cf;
*<**Suj**:cln|sn,**Obl**:pour-sn>;*
cat=nc;@SujNhum;@OblN-hum;@OblNhum;
lightverb=avoir|ressentir|éprouver;
%default

considération N_aa_18 *nc-2f*
100;Lemma;cf;
*<**Suj**:cln|sn,**Objde**:de-sn>;*
cat=nc;@SujNhum;
lightverb=avoir;
%default

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

- Reste à convertir : **expressions figées** et **adverbes**

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

L'analyseur syntaxique FRMG

FRMG est un analyseur **TAG** du français :

- issu de la compilation d'une **méta-grammaire**
- très **compacte** grâce à la factorisation des arbres
  (1 986 300 arbres défactorisés $\rightarrow \sim$ 75 630 arbres)
- exploitant les fonctionnalités de **DyALog**
  $\rightarrow$ environnement de programmation en logique (langage, compilateur, machine virtuel)

FRMG s'intègre dans une chaîne de traitement :

- en amont, avec **SXPipe** et le lexique **Lefff**
  - **SXPipe** : segmentation, token, corrections, entités nommées
    $\rightarrow$ retourne un **DAG** (graphes de mots)
  - **Lefff** : lexique morphosyntaxique et syntaxique du français
    $\rightarrow$ lien lexique/grammaire : ancrage par **hypertag**
- en aval, avec un module de **désambiguïsation** (basé sur des heuristiques)

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

Intégration des verbes [Tolone & Sagot 2011] et des noms prédicatifs

Intégration dans l'analyseur syntaxique FRMG (2)

Il faut donc :

- remplacer les entrées verbales du Lefff par $LGLex_{\text{Lefff}}$
- ajouter les entrées nominale de $LGLex_{\text{Lefff}}$
- 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, notée 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 Passage [Hamon *et al.* 2008]
  - 270 millions de mots : styles variés (encyclopédique, littéraire, journalistique, juridique, oral, etc.)
- métriques utilisées : celles définies à l'occasion de la campagne EASy d'évaluation des analyseurs syntaxiques [Paroubek *et al.* 2006]
  - évaluation en *chunks* et en *relations* ($\sim$ dépendances entre mots pleins)

Précautions

Avant d'interpréter les résultats de FRMG_{LGLex}

- FRMG utilise les entrées converties à partir des tables, et non pas les entrées telles qu'elles sont dans les tables
  - le processus de conversion décrit ici et son implémentation contiennent certainement des erreurs
- la majorité des noms prédicatifs n'ont pas pu être évalués car FRMG ne prend que ceux n'ayant pas de déterminant
- Passage 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 et Passage, contrairement aux tables
- $LGLex_{\text{Lefff}}$ ne contient pas toutes les entrées verbales nécessaires, et il a fallu en rajouter
  - il en manque encore puisque tous les verbes ne sont pas codés

Résultats

Résultats comparatifs Passage 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 Passage

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 au LIGM (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 (par exemple, le grec)

Références

- [Boons et al. 1976a] Boons J.-P., Guillet A. et Leclère C. *La structure des phrases simples en français : Classes de constructions transitives.* Rapport interne, LADL, CNRS, Paris 7. 1976.
- [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.
- [Gross 1986] Maurice Gross. *Grammaire transformationnelle du français : Syntaxe de l'adverbe*, volume 3. Paris : ASSTRIL. 1986.
- [Hamon et al. 2008] Olivier Hamon, Djamel Mostefa, Christelle Ayache, Patrick Paroubek, Anne Vilnat et Éric de La Clergerie. Passage : from French Parser Evaluation to Large Sized Treebank. Actes de LREC'08. Marrakech, Maroc. 2008.

Références (2)

- [Molinier & Levrier 2000] Molinier C. et Levrier F. *Grammaire des adverbes : description des formes en -ment.* Genève : Droz. 2000.
- [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.
- [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, pages 41-60. Hermès. Paris, France. 2008.
- [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.

Références (3)

- [Thomasset & de La Clergerie 2005] François Thomasset et 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'09, 8 pages. Avignon, France. 2009.
- [Tolone 2010] Elsa Tolone. Constructions définitoires des tables du Lexique-Grammaire : Le cas des complétives et des infinitives. Actes de MajecSTIC'10 (session poster), 4 pages. Bordeaux, France. 2010.
- [Tolone et al. 2010] Elsa Tolone, Stavroula Voyatzi et Christian Leclère. Constructions définitoires des tables du Lexique-Grammaire. Actes de LGC'10, pages 321-331. Belgrade, Serbie. 2010.
- [Tolone & Sagot 2011] Elsa Tolone et Benoît Sagot. Using Lexicon-Grammar tables for French verbs in a large-coverage parser. LNAI. Springer Verlag. 2011. À paraître.

Thèse

- http://www-igm.univ-mlv.fr/~tolone/phd.pdf