

Comment améliorer une chaîne de traitement syntaxique ?

Éric de la Clergerie
avec des contributions de

Benoît Sagot, Lionel Nicolas, Milagros Fernandez, Marie-Laure Guénot

<http://alpage.inria.fr>



Séminaire IGM
Marne-La-Vallée, 11 Janvier 2010

- 1 Contexte
- 2 Premières améliorations
- 3 Fouille d'erreurs
- 4 Suggérer des corrections d'erreurs
- 5 Fouille d'erreurs et apprentissage (non supervisé)

FRMG est un analyseur TAG du français

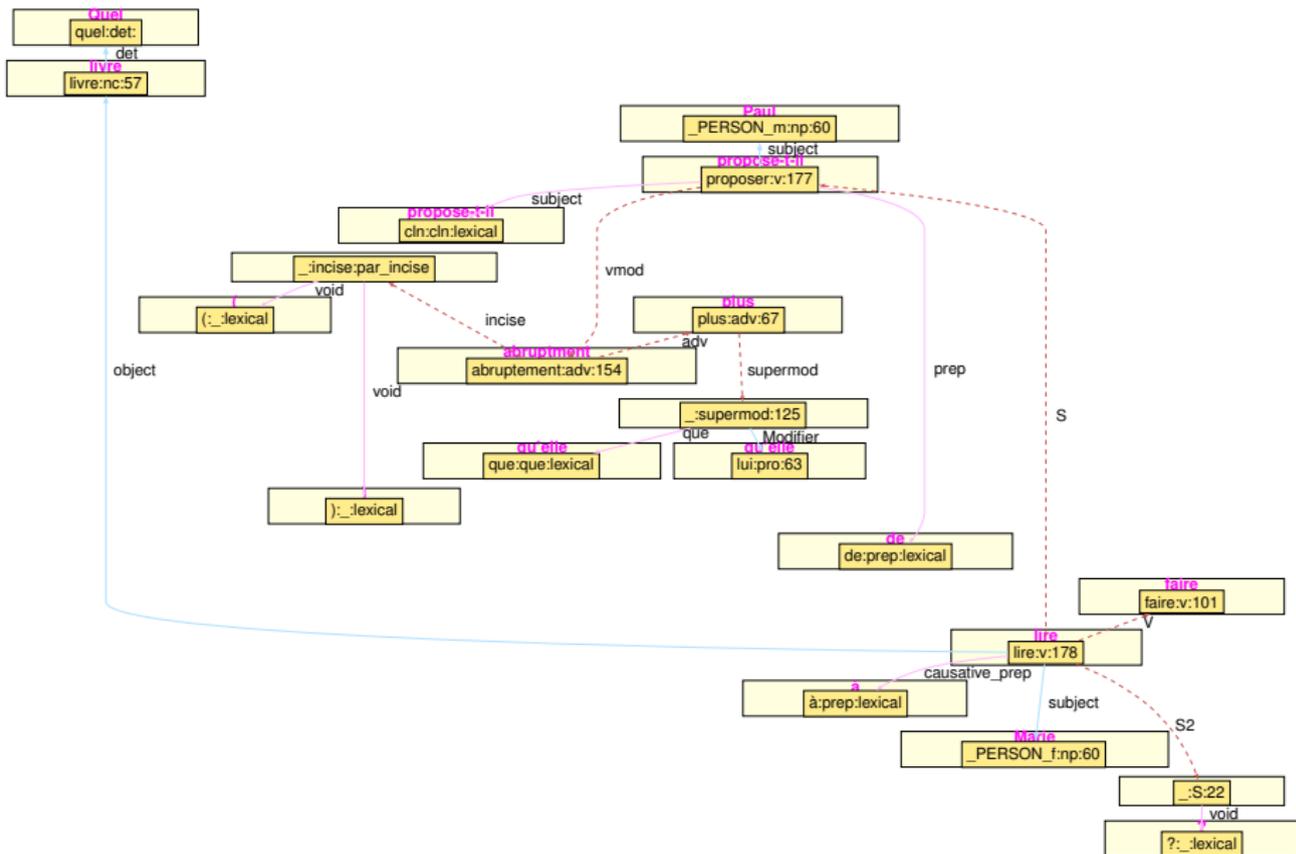
- issu de la compilation d'une méta-grammaire
- très compacte grâce à la factorisation des arbres
- exploitant les fonctionnalités de **DYALOG** environnement de programmation en logique (langage, compilateur, machine virtuel)

FRMG s'intègre dans une chaîne de traitement

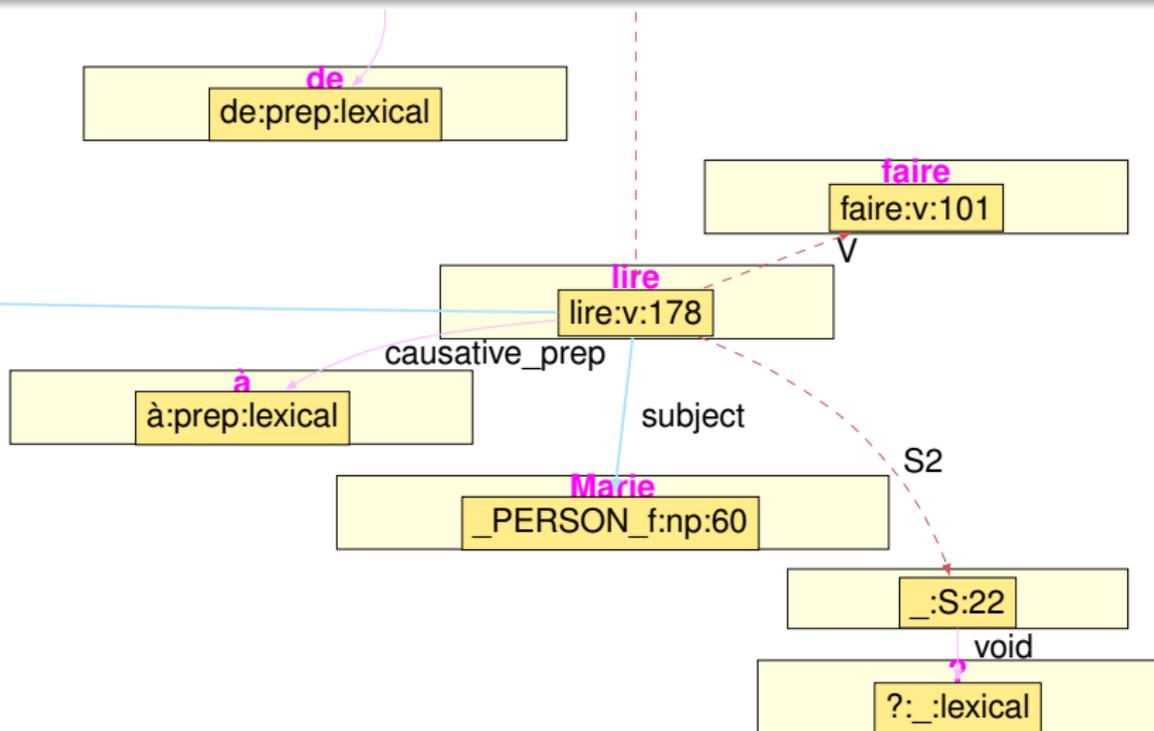
- en amont, avec **SXPIPE** et le lexique **LEFFF**
 - ▶ **SXPIPE**: segmentation, token, corrections, entités nommées retourne un DAG (treillis de mots)
 - ▶ **LEFFF** : lexique morphosyntaxique et syntaxique du français
- en aval, avec un module de désambiguïsation

FRMG et ses compagnons soumis à un processus de retour pour améliorer les performances.

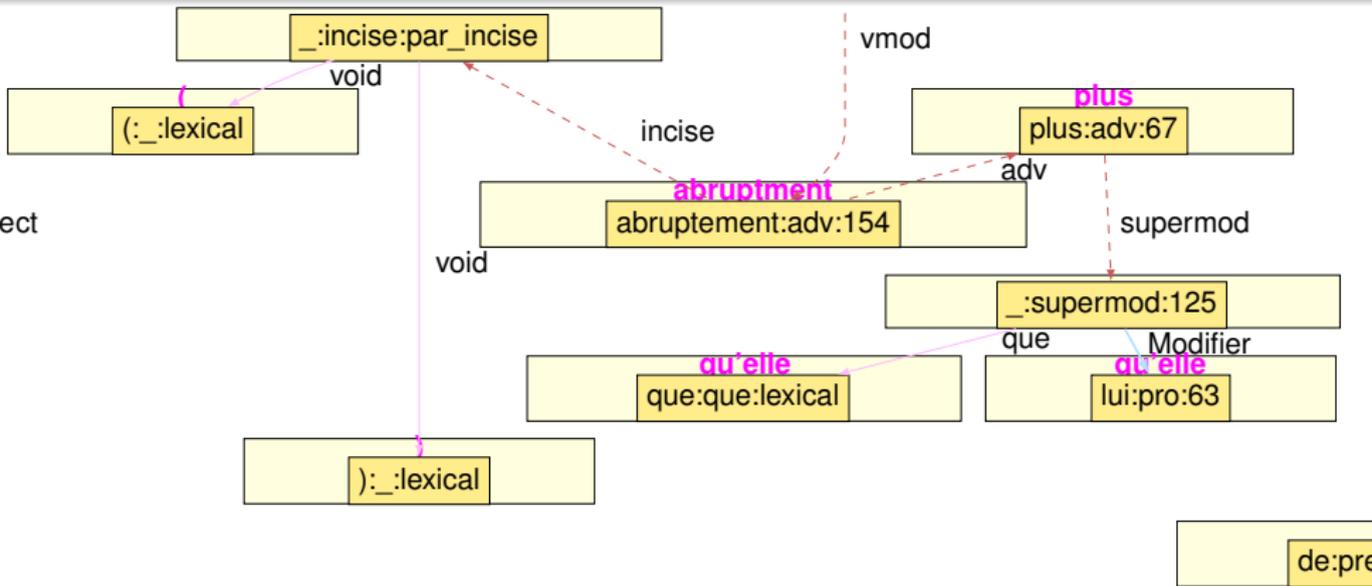
Quel livre Paul propose-t-il (plus abruptement qu'elle) de faire lire à Marie ?



Quel livre Paul propose-t-il (plus abruptement qu'elle) de faire lire à Marie ?



Quel livre Paul propose-t-il (plus abruptement qu'elle) de faire lire à Marie ?



Développement du lexique LEFF [Clément, Sagot]

- plus de 400 000 formes avec la distribution suivante de lemmes (en 2005):

verbes	noms communs	noms propres	adj	adv
6788	37183	52938	10024	2127

- Morphologie des verbes **automatiquement apprises** sur corpus (+ validation manuelle)
- Information syntaxique sur les verbes (sous-catégorisation, contrôle, ...) **promet** (*promises*)

```
v [pred='promettre_1<subj|ssubj|vsubj, (obj|scomp), (à-obj) >', cat=v, @SCompInd, @P3s]
v [...]
```

Grammaire hypertag #111

arg0	arg0	<table border="0"> <tr><td>extracted</td><td>-</td></tr> <tr><td>kind</td><td>subj</td></tr> <tr><td>pcas</td><td>-</td></tr> <tr><td>real</td><td>real0 - CS N2 PP S cln prel pri </td></tr> </table>	extracted	-	kind	subj	pcas	-	real	real0 - CS N2 PP S cln prel pri
extracted	-									
kind	subj									
pcas	-									
real	real0 - CS N2 PP S cln prel pri									
arg1	arg1	<table border="0"> <tr><td>extracted</td><td>-</td></tr> <tr><td>kind</td><td>kind1 - acomp obj prepacomp prepobj </td></tr> <tr><td>pcas</td><td>pcas1 + - apres à avec de par ...</td></tr> <tr><td>real</td><td>real1 - CS N N2 PP S V adj cla ...</td></tr> </table>	extracted	-	kind	kind1 - acomp obj prepacomp prepobj	pcas	pcas1 + - apres à avec de par ...	real	real1 - CS N N2 PP S V adj cla ...
extracted	-									
kind	kind1 - acomp obj prepacomp prepobj									
pcas	pcas1 + - apres à avec de par ...									
real	real1 - CS N N2 PP S V adj cla ...									
arg2	arg2	<table border="0"> <tr><td>extracted</td><td>-</td></tr> <tr><td>kind</td><td>kind2 - prepacomp prepobj prepscomp prepvcomp scomp vcomp wh-comp </td></tr> <tr><td>pcas</td><td>pcas2 - + apres à ...</td></tr> <tr><td>real</td><td>real2 - CS N N2 PP S ...</td></tr> </table>	extracted	-	kind	kind2 - prepacomp prepobj prepscomp prepvcomp scomp vcomp wh-comp	pcas	pcas2 - + apres à ...	real	real2 - CS N N2 PP S ...
extracted	-									
kind	kind2 - prepacomp prepobj prepscomp prepvcomp scomp vcomp wh-comp									
pcas	pcas2 - + apres à ...									
real	real2 - CS N N2 PP S ...									
cat	v									
diathesis	active									
refl	refl									

Grammaire hypertag #111

arg0	arg0	[extracted	-]
			kind	subj	
			pcas	-	
			real	real0 - CS N2 PP S cln prel pri	
arg1	arg1	[extracted	-]
			kind	kind1 - acomp obj prepacomp prepobj	
			pcas	pcas1 + - apres à avec de par ...	
			real	real1 - CS N N2 PP S V adj cla ...	
arg2	arg2	[extracted	-]
			kind	kind2 - prepacomp prepobj prepscomp prepvcomp scomp vcomp wh-comp	
			pcas	pcas2 - + apres à ...	
			real	real2 - CS N N2 PP S ...	
cat			v		
diathesis			active		
refl	refl				

Lexique hypertag «promettre»

arg0	[kind	subj -]
		pcas	-	
arg1	[kind	obj scomp -]
		pcas	-	
arg2	[kind	prepobj -]
		pcas	à -	
refl			-	

Grammaire hypertag #111

Lexique hypertag «promettre»

arg0	<code>arg0</code>	extracted - kind subj pcas - real <code>real0</code> - CS N2 PP S cln prel pri
arg1	<code>arg1</code>	extracted - kind <code>kind1</code> - acomp obj prepacomp prepobj pcas <code>pcas1</code> + - apres à avec de par ... real <code>real1</code> - CS N N2 PP S V adj cla ...
arg2	<code>arg2</code>	extracted - kind <code>kind2</code> - prepacomp prepobj prepscomp prepvcomp scomp vcomp whcomp pcas <code>pcas2</code> - + apres à ... real <code>real2</code> - CS N N2 PP S ...
cat	v	
diathesis	active	
refl	<code>refl</code>	-

arg0	[kind subj -] pcas -
arg1	[kind obj scomp -] pcas -
arg2	[kind prepobj -] pcas à -
refl	-

FRMG retourne l'ensemble des analyses sous forme des *forêts partagées de dérivations*:

- chaque étape de dérivation (subst, adj) applique un arbre τ_1 sur un noeud N d'un autre arbre τ_2

Conversion en une forêt partagée de dépendances

- arcs de la forme $\text{anchor}(\tau_1) \longrightarrow_N \text{anchor}(\tau_2)$
- introduction de pseudo-ancres vides pour les arbres non-lexicalisés

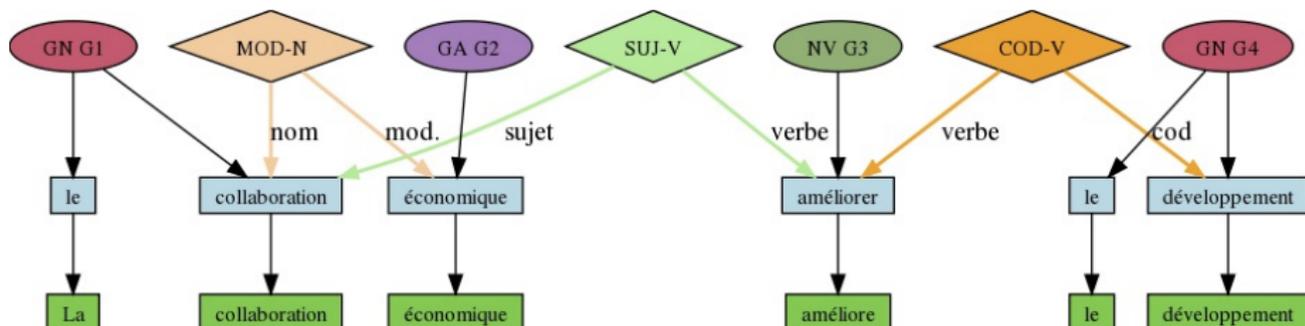
Représentation DEP XML s'appuyant sur

- *cluster* pour les formes
- *node* pour lemmes, pos, arbres ancrés, dérivations associées
- *edge* liant les nodes, associés à un sous-ensemble des dérivation *deriv* du noeud gouverneur
- d'autres informations dont constituants *op*

- algorithme de type 1-best en programmation dynamique, écrit en **DYALOG**
- sommes de poids sur les dépendances
- poids fournis par des règles portant sur la dépendance et ses voisines
- poids manuellement définis
quelques tentatives d'apprentissage
- temps de traitement du même ordre que pour l'analyse

```
%% Penalize inverted subjects (especially for robust mode)
edge_cost_elem( Name:: '-INVERTED_SUBJ',
edge{ label => subject,
      source => node{ cluster => cluster{ right => R } },
      target => node{ cluster => cluster{ left => L } }
},
W
) :-
rule_weight(Name,W,-1000),
R =< L
.
```

- vers formats EASy et Passage:
6 types de *chunks* et 14 types de dépendance
- plus “superficiel” que sorties FRMG
- erreurs dues à la conversion



Jan. 2004

Lexique MorphoLEFFF

Dec. 2004
Campagne Easy

Jan. 2004

Lexique Morpho **LEFFF**

Dec. 2004

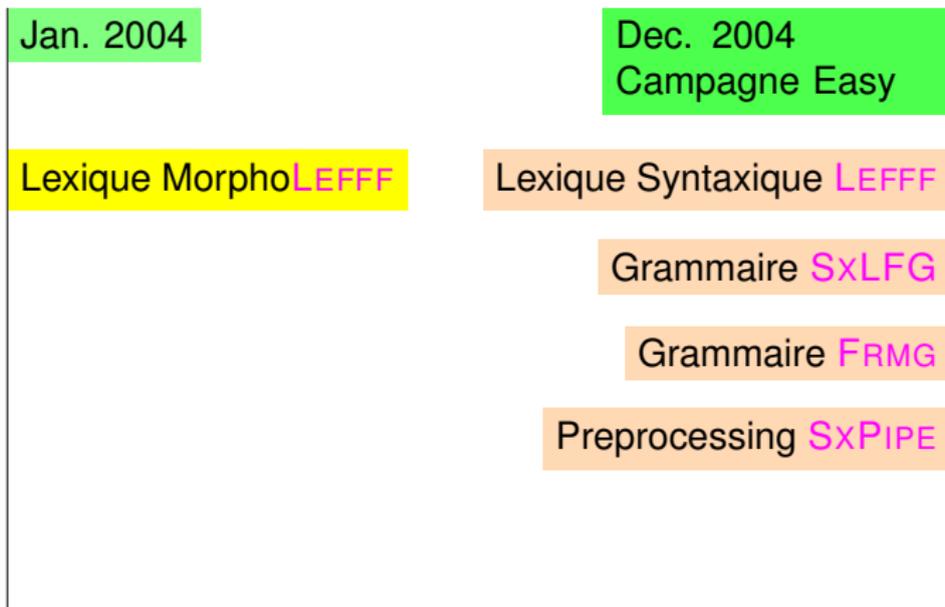
Campagne Easy

Lexique Syntaxique **LEFFF**

Grammaire **SXLFG**

Grammaire **FRMG**

Preprocessing **SXPIPE**



Résultats corrects pour EASy, mais ressources à améliorer !
Depuis: campagnes Passage en 2007 et en 2009 (100Mmots)

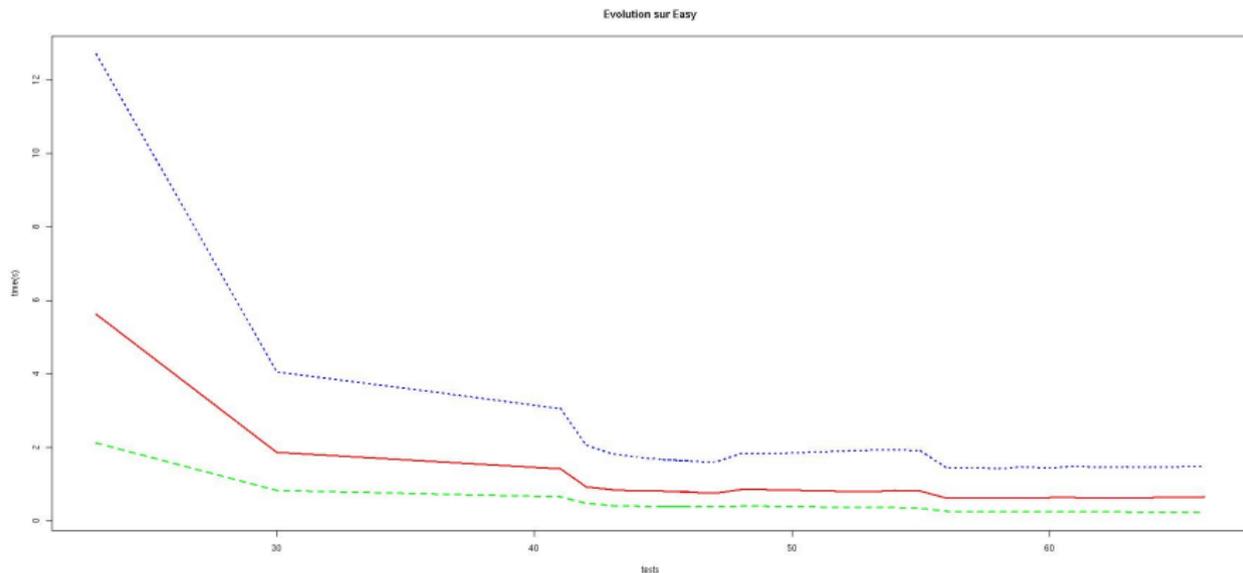
- 1 Contexte
- 2 Premières améliorations**
- 3 Fouille d'erreurs
- 4 Suggérer des corrections d'erreurs
- 5 Fouille d'erreurs et apprentissage (non supervisé)

Cela porte sur 4 axes

- augmenter la **couverture** en terme d'analyses complètes
 - ⇒ méthodes non-supervisées sur large corpus: **fouille d'erreurs**
 - ⇒ ouverture vers des techniques d'apprentissage
- améliorer la **qualité** des analyses
 - ⇒ méthodes supervisées: utilisation d'une référence
- réduire le taux d'**ambiguïté** des analyses
- réduire les **temps d'analyse** et de désambiguïsation.
important mais pas essentiel: utilisation grille de machines (GRID5000)

Ces axes sont en partie contradictoires

Évolutions vitesse FRMG (2008)



Easydev: 3868 phrases, 58.20% couverture, timeout (20s): 0.4%

- Multiples runs sur jeux de tests et corpus, avec statistiques
- Test de nombreuses optimisations, la plupart inefficaces
- Mais gains instables: variations importantes en fonction de la grammaire

- Améliorer la qualité en exploitant des données de référence (treebank) données EASy
- résultats globaux, par type de corpus, groupes et relations log, évaluations, matrices de confusion
- visualisation des résultats
Gestionnaire d'annotations avec service WEB **EASYREF**

Matrice de confusion

réf \implies hyp	B_GN	GN	E_GN	BE_GN	B_GP
B_GN	5459 (91%)	52 (0.88%)	14 (0.28%)	115 (1.94%)	118 (1.99%)
GN	50 (4.65%)	725 (67%)	149 (13.86%)	19 (1.77%)	12 (1.12%)
E_GN	4 (0.07%)	68 (1.15%)	5345 (90.04%)	140 (2.36%)	10 (0.17%)
BE_GN	106 (3.22%)	45 (1.37%)	78 (2.37%)	2663 (80.97%)	7 (0.21%)
B_GP	166 (2.02%)	30 (0.37%)	2 (0.02%)	30 (0.37%)	7760 (94.61%)

Nouveau: Matrices de différences entre runs.

Évolution des performances

	% analyse	Groupes			Relations		
	totales	prec.	rappel	f	prec.	rappel	f
R6 (05/07)	42.16%	78.12%	71.27%	74.54%	62.29%	46.63%	53.34%
R27 (12/07)	56.06%	83.66%	82.90%	83.28%	64.27%	55.66%	59.65%
R76 (02/09)	59.47%	84.23%	82.91%	85.56%	63.36%	55.62%	59.24%
R101 (06/09)	59.56%	83.24%	79.63%	81.40%	63.1%	53.40%	57.85%
R139 (09/09)	64.73%	87.41%	86.00%	86.70%	65.10%	59.03%	61.92%
R157 (10/09)	67.03%	87.71%	86.84%	87.28%	65.62%	60.26%	62.82%
R206 (01/10)	65.79%	87.92%	88.60%	88.26%	66.12%	62.13%	64.06%

Campagne	f-mesure groupes	f-mesure relations
2004	69%	41%
2007	89%	63%

Note: Nos outils d'évaluation donnent des valeurs plus faibles que les valeurs officielles.

Jeux	#phrases	Couv.	t. moy. (s)	amb.	couv. 09/09
EUROTRA	334	100%	0.15	0.63	
TSNLP	1161	95.07%	0.07	0.46	
EasyDev	3879	64.73%	0.93	1.04	
JRCacquis	1.1M	51.26%	1.41	1.1	59.46%
Europarl	0.8M	70.19%	1.69	1.36	78.33%
EstRep	1.6M	67.05%	0.69	0.92	75.06%
Wikipedia	2.2M	69.11%	0.49	0.87	79.48%
Wikisource	1.5M	61.08%	0.71	0.89	66.79%
AFP	1.6M	52.15%	0.51	1.06	

- Pour améliorer la couverture, recherche des manques dans le lexique, la grammaire, ...
- traitement de gros corpus suivi de fouille d'erreurs
 - ▶ identifier les mots trop souvent présents dans des phrases non analysables
 - ▶ surtout si en co-occurrence avec des mots sans problèmes
 - ▶ processus itératif de type EM
 - ▶ fournit des phrases où le mot est le principal suspect
- possibilité de suggérer des corrections:
 - ▶ ré-analyse les phrases en sous-spécifiant le suspect
 - ▶ fait ressortir les analyses les plus fréquentes devenues possibles au niveau du suspect

- Mots manquants
- Catégorie manquante pour une forme telle que nom \rightsquigarrow adjectif (et réciproquement)
- Mauvais auxiliaire pour un participe passé
- Cadre incomplet de sous-catégorisation
 - ▶ arguments manquants
 - ▶ arguments non marqués comme optionnels
 - ▶ réalisations manquantes
 - objet nominal \rightsquigarrow objet phrastique (et réciproquement)
- Mauvaise info morpho-syntaxique: genre, nombre, . . .

Trouver les erreurs à la main !



www.jolyon.co.uk

Fouille d'erreurs



Comment le faire ?

- Repose sur du retour (feedback) d'analyse de corpus

Comment le faire ?

- Repose sur du retour (feedback) d'analyse de corpus
- **Question:** Comment minimiser le coût humain d'exploitation du feedback
⇒

Comment le faire ?

- Repose sur du retour (feedback) d'analyse de corpus
- **Question:** Comment minimiser le coût humain d'exploitation du feedback
⇒
 - ▶ bon modèle statistique

Comment le faire ?

- Repose sur du retour (feedback) d'analyse de corpus
- **Question:** Comment minimiser le coût humain d'exploitation du feedback
 - ⇒
 - ▶ bon modèle statistique
 - ▶ bonnes interfaces

Comment le faire ?

- Repose sur du retour (feedback) d'analyse de corpus
- **Question:** Comment minimiser le coût humain d'exploitation du feedback
⇒
 - ▶ bon modèle statistique
 - ▶ bonnes interfaces
 - ▶ bonne intégration de toutes les sources d'information (corpus, grammaire, lexique, chaîne de pré-traitement, ...)

Comment le faire ?

- Repose sur du retour (feedback) d'analyse de corpus
- **Question:** Comment minimiser le coût humain d'exploitation du feedback
⇒
 - ▶ bon modèle statistique
 - ▶ bonnes interfaces
 - ▶ bonne intégration de toutes les sources d'information (corpus, grammaire, lexique, chaîne de pré-traitement, ...)
 - ▶ reproductibilité (en gardant la trace de ce qui est fait)

Comment le faire ?

- Repose sur du retour (feedback) d'analyse de corpus
- **Question:** Comment minimiser le coût humain d'exploitation du feedback

⇒

- ▶ bon modèle statistique
- ▶ bonnes interfaces
- ▶ bonne intégration de toutes les sources d'information (corpus, grammaire, lexique, chaîne de pré-traitement, ...)
- ▶ reproductibilité (en gardant la trace de ce qui est fait)
- ▶ un maximum d'interprétation réalisé par programmation

Comment le faire ?

- Repose sur du retour (feedback) d'analyse de corpus
- **Question:** Comment minimiser le coût humain d'exploitation du feedback
⇒
 - ▶ bon modèle statistique
 - ▶ bonnes interfaces
 - ▶ bonne intégration de toutes les sources d'information (corpus, grammaire, lexique, chaîne de pré-traitement, ...)
 - ▶ reproductibilité (en gardant la trace de ce qui est fait)
 - ▶ un maximum d'interprétation réalisé par programmation
- ⇒ expérience (inspirée de [Van Noord](#))

Comment le faire ?

- Repose sur du retour (feedback) d'analyse de corpus
- **Question:** Comment minimiser le coût humain d'exploitation du feedback
⇒
 - ▶ bon modèle statistique
 - ▶ bonnes interfaces
 - ▶ bonne intégration de toutes les sources d'information (corpus, grammaire, lexique, chaîne de pré-traitement, ...)
 - ▶ reproductibilité (en gardant la trace de ce qui est fait)
 - ▶ un maximum d'interprétation réalisé par programmation
- ⇒ expérience (inspirée de **Van Noord**)
 - ▶ utilisation de phrases non-analysables pour détecter les entrées lexicales erronées et les constructions syntaxiques manquantes.

Comment le faire ?

- Repose sur du retour (feedback) d'analyse de corpus
- **Question:** Comment minimiser le coût humain d'exploitation du feedback
⇒
 - ▶ bon modèle statistique
 - ▶ bonnes interfaces
 - ▶ bonne intégration de toutes les sources d'information (corpus, grammaire, lexique, chaîne de pré-traitement, ...)
 - ▶ reproductibilité (en gardant la trace de ce qui est fait)
 - ▶ un maximum d'interprétation réalisé par programmation
- ⇒ expérience (inspirée de **Van Noord**)
 - ▶ utilisation de phrases non-analysables pour détecter les entrées lexicales erronées et les constructions syntaxiques manquantes.
 - ▶ combiner plusieurs analyseurs pour mieux cerner les sources d'erreurs

Que peut-on améliorer ?

Toute ressource linguistique de grande taille et/ou complexe

- Lexique
- Grammaire
- mais aussi des corpus bruts ou annotés ([treebank](#))

- 1 Contexte
- 2 Premières améliorations
- 3 Fouille d'erreurs**
- 4 Suggérer des corrections d'erreurs
- 5 Fouille d'erreurs et apprentissage (non supervisé)

Trouver les formes sur-représentées dans des phrases non-analysables

$$\text{err}(f) = \frac{\# \text{failed occurrences } f}{\# \text{occurrences } f}$$

Trouver les formes sur-représentées dans des phrases non-analysables

$$\text{err}(f) = \frac{\# \text{failed occurrences } f}{\# \text{occurrences } f}$$

- OK pour des ressources déjà de bonne qualité
- certains mots fréquents peuvent être en cooccurrence avec beaucoup de formes erronées
mots fonctionnels, ponctuations, ...
- n'indique pas vraiment qu'une forme est la cause d'un échec d'analyse
⇒ difficile de fournir de bonnes phrases d'exemple

*Trouver les **formes suspectes** qui apparaissent plus souvent qu'attendu dans des phrases non analysées, en co-occurrence avec des mots des formes qui tendent à ne pas être **suspectes**.*

*Trouver les **formes suspectes** qui apparaissent plus souvent qu'attendu dans des phrases non analysées, en co-occurrence avec des mots des formes qui tendent à ne pas être **suspectes**.*

- définition récursive de la notion de **suspect**
⇒ itération jusqu'à obtention d'un point fixe
- identification des suspects dans les phrases
⇒ bon échantillon de (courtes) phrases échec pour un suspect donné.

Le modèle statistique

- Corpus comme ensemble de phrases s_i , composées d'occurrences $o_{i,j}$ associées à des formes $F(o_{i,j})$.

Le modèle statistique

- Corpus comme ensemble de phrases s_i , composées d'occurrences $o_{i,j}$ associées à des formes $F(o_{i,j})$.
- Taux moyen de suspicion \bar{S} pour 1 occ
= chance qu'une occurrence soit une cause d'erreur

$$\bar{S} = \frac{\sum_i \text{error}(s_i)}{\text{OCC}_{\text{total}}}$$

- Corpus comme ensemble de phrases s_i , composées d'occurrences $o_{i,j}$ associées à des formes $F(o_{i,j})$.
- Taux moyen de suspicion \bar{S} pour 1 occ
= chance qu'une occurrence soit une cause d'erreur

$$\bar{S} = \frac{\sum_i \text{error}(s_i)}{\text{OCC}_{\text{total}}}$$

- Taux local de suspicion $S_{i,j}$ pour une occurrence donnée $o_{i,j}$
($S_{i,j} = 0$ pour une phrase succès, $S_{i,j} = x$ pour une phrase échec)

- Corpus comme ensemble de phrases s_i , composées d'occurrences $o_{i,j}$ associées à des formes $F(o_{i,j})$.
- Taux moyen de suspicion \bar{S} pour 1 occ
= chance qu'une occurrence soit une cause d'erreur

$$\bar{S} = \frac{\sum_i \text{error}(s_i)}{\text{OCC}_{\text{total}}}$$

- Taux local de suspicion $S_{i,j}$ pour une occurrence donnée $o_{i,j}$
($S_{i,j} = 0$ pour une phrase succès, $S_{i,j} = x$ pour une phrase échec)
- Taux moyen global de suspicion S_f pour une forme f

$$S_f = \frac{1}{|\mathcal{O}_f|} \cdot \sum_{o_{i,j} \in \mathcal{O}_f} S_{i,j} \quad \text{with} \quad S_{i,j} = \text{error}(s_i) \cdot \frac{S_{F(o_{i,j})}}{\sum_{1 \leq j \leq |s_i|} S_{F(o_{i,j})}}$$

- Corpus comme ensemble de phrases s_i , composées d'occurrences $o_{i,j}$ associées à des formes $F(o_{i,j})$.
- Taux moyen de suspicion \bar{S} pour 1 occ
= chance qu'une occurrence soit une cause d'erreur

$$\bar{S} = \frac{\sum_i \text{error}(s_i)}{\text{OCC}_{\text{total}}}$$

- Taux local de suspicion $S_{i,j}$ pour une occurrence donnée $o_{i,j}$
($S_{i,j} = 0$ pour une phrase succès, $S_{i,j} = x$ pour une phrase échec)
- Taux moyen global de suspicion S_f pour une forme f

$$S_f = \frac{1}{|\mathcal{O}_f|} \cdot \sum_{o_{i,j} \in \mathcal{O}_f} S_{i,j} \quad \text{with} \quad S_{i,j} = \text{error}(s_i) \cdot \frac{S_{F(o_{i,j})}}{\sum_{1 \leq j \leq |s_i|} S_{F(o_{i,j})}}$$

- \Rightarrow interactions entre le niveau global (corpus) et le niveau local (phrase)
 \Rightarrow recherche d'un point-fixe!

- calcul itératif

$$S_{i,j}^{(0)} = \text{error}(s_i) / |s_i|$$

$$S_f^{(n+1)} = \frac{1}{|\mathcal{O}_f|} \cdot \sum_{o_{i,j} \in \mathcal{O}_f} S_{i,j}^{(n)}$$

$$S_{i,j}^{(n+1)} = \text{error}(s_i) \cdot \frac{S_{F(o_{i,j})}^{(n+1)}}{\sum_{1 \leq j \leq |s_i|} S_{F(o_{i,j})}^{(n+1)}}$$

Motivation: lissage du modèle pour les formes à faible fréquence mais les expériences pratiques ne valident pas l'intérêt du lissage

lissage obtenu par calcul d'un barycentre, en assumant que le taux de suspicion pour une forme peu fréquent doit être proche du taux de suspicion de base \bar{S} .

$$\tilde{S}_f^{(n)} = \lambda(|\mathcal{O}_f|) \cdot S_f^{(n)} + (1 - \lambda(|\mathcal{O}_f|)) \cdot \bar{S}$$

avec $\begin{cases} \lambda(|\mathcal{O}_f|) = 1 - e^{-\beta|\mathcal{O}_f|} \\ \beta = 0.1 \end{cases}$

L'ordonnement des formes obtenu au travers de S_f , mais en focalisant en priorité sur les formes fréquentes (principe d'efficacité):

$$M_f = S_f \cdot \ln |\mathcal{O}_f|$$

Cet ordonnancement peut expliquer la faible utilité du lissage

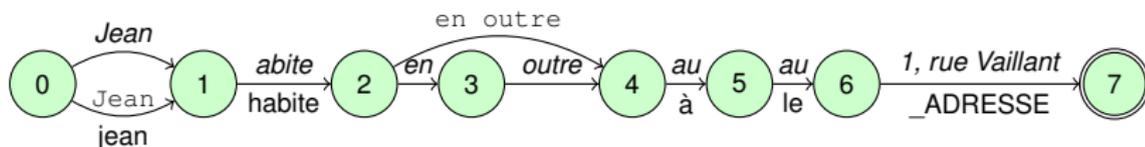
- En PERL, en utilisant des structures partagées pour des accès rapides aux information:
 - ▶ une table des formes, avec pour chaque forme f , une structure $s(f)$.
 - ▶ une table des phrases, avec pour chaque phrase s_i , une structure $p(i)$ contenant une liste des occurrences avec accès direct aux entrées $s(F(o_{ij}))$
- $s(f)$ maintient une **liste de phrases** où f est un des principaux suspects (c-a-d avec un fort taux local de suspicion)

- Lexique **LEFF**

plus de 400 000 formes avec la distribution suivante de lemmes:

verbes	noms communs	noms propres	adj	adv
6788	37183	52938	10024	2127

- Chaîne de pré-traitement syntaxique **SXPIPE**



2 analyseurs:

- **FRMG** (Meta-Grammaire \rightsquigarrow hybride TAG/TIG / **DYALOG**)
 - ▶ analyses complètes ou analyses partielles
- **SXLFG** (LFG/**SYNTAX**)

Note: Deux analyseurs distincts mais avec le même lexique **LEFFF**

2 corpus: Monde Diplomatique et EASy

corpus	#sentences	#success (%)	#forms	#occ	\bar{S} (%)	Date
MD/FRMG	330 938	136 885 (41.30%)	255 616	10 422 926	1.86%	Jul. 05
MD/SxLFG	630 000	337 437 (53.56%)	373 986	15 840 364	1.85%	Dec. 05
EASy/FRMG	39 872	16 477 (41.32%)	61 135	878 156	2.66%	Dec. 05
EASy/SxLFG	39 872	21 067 (52.84%)	61 135	878 156	2.15%	Dec. 05

Pour MD/FRMG:

- 200 itérations (~ 2700 s sur PC 3.2GHz avec 1 Go RAM)
- $\Rightarrow 18\,492$ formes *pertinentes*
avec filtrage $S_f^{(200)} > 1,5 \cdot \bar{S}$ et $|O_f| > 5$.
- convergence en pratique: variation moyenne de moins de 0.01% sur les 1000 premières formes
- répartition des suspects:

#occ	> 100 000	> 10 000	> 1000	> 100	> 10
# formes	13	84	947	8345	40 393
# suspects	1	13	177	1919	12 022
%	7.6%	15.5%	18.7%	23%	29.8%

Interface WEB avec support javascript (coté client) et CGI + DB (coté serveur).

Browsing errors for results5 [iter=200]

- 27 voilà
- 28 lui-même
- 29 jusque
- 30 emparé
- 31 p.
- 32 endettés
- 33 il est vrai que /il est vrai que
- 34 demeure
- 35 -
- 36 azimuts
- 37 50 /à
- 38 rase
- 39 the /rhé
- 40 dus
- 41 eux-mêmes
- 42 elle-même
- 43 coopérer
- 44 notamment
- 45 soucie
- 46 demeurerait
- 47 monsieur
- 48 censé
- 49 autorisée
- 50 censée
- 51 quant aux /quant à
- 52 rend
- 53 censés
- 54 quoi
- 55 taliban
- 56 disputent
- 57 prospères
- 58 d'en bas /en bas
- 59 endetté
- 60 qu'a / uw
- 61 Et /et

Enter rank (or start:end:key) [Mail this page](#)

[edit comment](#)

manque la construction attributive (demeurer<subj,acomp>)

Statistical info on **demeure/demeure**

rank	#occ.	#failed	%failed weight	%failed sentences	orate
34	870	706	24.64%	81.15%	7.27

history:

#iteration	200	199	195	185	175	165	155	145	135	125	115	105	95	90
weight	24.64%	24.64%	24.64%	24.65%	24.65%	24.66%	24.68%	24.69%	24.71%	24.73%	24.75%	24.78%	24.81%	24.83%

Lefff info for **demeure**

```
nc [pred='demeure_____1<(subj),(de-obj),(de-vcomp|à-vcomp)>',cat=nc,#f]
v [pred='demeurer_____1<subj>',cat=v,#imperative,#Y2s]
v [pred='demeurer_____1<subj>',cat=v,#P813a]
```

Failed sentences with **demeure/demeure** as most probable cause for failure

- [mondediplo_01#19948] L'armée **demeure** une force majeure
- [mondediplo_02#22126] LE FN **demeure** l'unique parti à défendre les négationnistes dans son programme .
- [mondediplo_04#7744] Le pétrole **demeure** l'enjeu principal .
- [mondediplo_01#19379] L'EUROPE **demeure** un projet à deux vitesses .
- [mondediplo_01#19984] Certes , l'Indonésie **demeure** la grande puissance régionale .
- [mondediplo_01#28830] L'histoire **demeure** cependant la principale discipline d'enseignement .
- [mondediplo_05#15643] En mer Rouge et dans la corne de l'Afrique , la situation **demeure** très incertaine .
- [mondediplo_02#17949] Le père **demeure** le chef exclusif de la famille .
- [mondediplo_04#10602] Une question toutefois **demeure** obscure .
- [mondediplo_06#19376] Le suédois **demeure** la deuxième langue officielle du pays .
- [mondediplo_06#20791] Quant à la Chine , elle **demeure** un grave sujet d'inquiétude .
- [mondediplo_06#31057] Or le social **demeure** une pièce rapportée de la construction européenne .
- [mondediplo_05#26084] Elle **demeure** nécessaire et enrichissante .

- lexique **LEFFF**: mauvaise ou incomplète sous-catégorisation, argument verbal optionnel, verbes pronominaux, ...
- Analyseurs (**FRMG** et **SXLFG**)
- Segmentation (**SXPIPE**):
ponctuations finales, entités nommées, marque de quotation, ...
- Corpus: certains segments ne sont pas vraiment des phrases !
Note: pas de nettoyage de corpus avant traitement

Résultats qualitatif (MD/FRMG)

Rang	Token(s)/forme	$S_f^{(200)}$	$ \mathcal{O}_f $	err(f)	M_f	cause d'erreur
1	;	65%	13379	100%	6.16	FRMG: « ; » ne peut terminer les phrases dans
2	*	83%	199	100%	4.40	corpus: phrases comportant le seul mot « * »
3	(...)	58%	1835	100%	4.37	SXPIPE: on devrait en faire un mot ignorable
4	coll/coll.	74%	257	93%	4.09	FRMG: ne sait pas analyser «L'Harmattan , col
5	jour	44%	3005	100%	3.49	LEFFF (ancienne version): manquait la forme
6	[...]	61%	173	100%	3.16	SXPIPE: on devrait en faire un mot ignorable
7	feu	47%	643	100%	3.01	LEFFF: « faire long feu » ; FRMG: adj. pré-déter
8	date	45%	749	100%	2.97	FRMG (ancienne version): pb sur les verbes su
9	confiance	44%	666	100%	2.89	FRMG(ancienne version): pb sur les verbes su
10	80/à	57%	138	100%	2.83	SXPIPE: bug (idem aux rangs 12: 70/à, 13: 60
11	etc/etc.	44%	493	100%	2.74	LEFFF: ce n'est pas une ponctuation...
15	voici	43%	269	93%	2.43	FRMG: grammaire incomplète
16	qu'elle/quelle	30%	3052	96%	2.37	SXPIPE (ancienne version): bug (id. 17:qu'elle
17	contrairement	39%	313	100%	2.23	LEFFF: Manque la notion de sous-cat adverbial
18	fiche	63%	34	94%	2.23	LEFFF: manque comme nom commun (!)
19	emparé	66%	29	100%	2.22	LEFFF: Manquent les constr. pronominales (s
20	demeurent	37%	398	87%	2.21	LEFFF: Manque la construction attributive
100	investir	30%	136	86%	1.48	LEFFF: l'objet direct n'est pas obligatoire
101	clôt	47%	23	96%	1.48	LEFFF: Manquent les constr. pronominales (s
102	demain	25%	378	81%	1.48	LEFFFet FRMG: Peut former un GN saturé (po
103	Seuls/seuls	29%	169	95%	1.48	FRMG: adj. pré-déterminant (id. au rang 104)
105	autoproclamé	50%	19	100%	1.47	LEFFF: Manque la construction attributive
106	renchérit	45%	25	92%	1.46	LEFFFet FRMG: Traiter les constructions narra
107	emparée	50%	18	100%	1.46	LEFFF: Manquent les constr. pronom. (id. rang
108	Mille/_NUMBER	56%	13	100%	1.45	SXPIPE: Mille et une pas reconnu comme un

Motivation: meilleur focus sur les erreurs lexicales, mettant de coté les erreurs grammaticales

Rang	Token(s)/forme		cause d'erreur
1	(...)	3,55	SXPIPE: en faire un mot que l'on peut ignorer (<code>_EPSILON</code>)
2	[...]	2,85	SXPIPE: comme pour <i>[...]</i>
3	demeurent	2,23	LEFFF: Manque la construction attributive
4	Premières/Premiere	2,09	SXPIPE: bug
5	emparé	2,05	LEFFF: Manquent les constr. pronominales (<i>s'emparer</i>)
6	endettés	2,04	LEFFF: Manque en tant qu'adjectif
7	lui-même	1,99	LEFFF: À mettre comme un adjectif spécial
8	endetté	1,87	LEFFF: comme pour <i>endettés</i>
9	elle-même	1,81	LEFFF: À mettre comme un adjectif spécial
10	d'en_bas/en_bas	1,79	LEFFF et grammaires: Constr. 'prep+adv' pour certains adv
11	larvée	1,78	LEFFF: Manque comme adjectif
12	-/-	1,75	grammaires: gestion du tiret parenthétique

Comparaison de méthodes (MD/SxLFG)

Rank	la notre		Van Noord		maxent	
	Token(s)/form	Eval	Token(s)/form	Eval	Token(s)/form	Eval
1	____/_ UNDERSCORE	++	*	+	pour	-
2	(...)	++	,	-)	-
3	2_]/_ NUMBER	++	livre	-	à	-
4	privées	++	.	-	qu'il/qu'	-
5	Haaretz/_Uw	++	de	-	sont	-
6	contesté	++	;	-	le	-
7	occupés	++	:	-	qu'un/qu'	+
8	privée	++	la	-	qu'un/un	+
9	[...]	++	étrangères	-	que	-
10	faudrait	++	lecteurs	-	pourrait	-

Utilisation de bigrammes (MD/SxLFG)

Rang	Tokens et formes	M_f	Cause
4	Toutes/toutes les	2.73	SXLFG: badly treated pre-determiner adjective
6	y en	2.34	SXLFG: problem with the construction <i>il y en a...</i>
7	in "	1.81	LEFFF: <i>in</i> misses as a preposition, which happens before book titles (hence)
10	donne à	1.44	LEFFF: <i>donner</i> should sub-categorize à-vcomps (<i>donner à voir...</i>)
11	de demain	1.19	LEFFF: <i>demain</i> misses as common noun (standard adv are not preceded b
16	(22/_NUMBER	0.86	SXLFG: footnote references not treated
16	22/_NUMBER)	0.86	SXLFG: as above

Bigrammes (et n -grams plus longs) utiles pour détecter:

- les expressions figées
- les constructions grammaticales manquantes

Nouvelle expérience (Juin 2008)

1 corpora: Monde Diplomatique

corpus	#sentences	#success (%)	#forms	#occ	S (%)	Date
MD/ FRMG	359 836	203 845 (56.65%)	231 637	7 703 339	2.02%	June. 08
MD/ FRMG	330 938	136 885 (41.30%)	255 616	10 422 926	1.86%	Jul. 05

10 059 pertinent forms (vs 18 492 en 2005)

répartition des suspects:

#occ	> 100 000	> 10 000	> 1000	> 100	> 10
# forms	16	62	564	5 384	24 655
# suspects	0	7	82	835	7 118
%	0.0%	11.3%	14.5%	15.5%	28.9%
# forms	13	84	947	8 345	40 393
# suspects	1	13	177	1 919	12 022
%	7.6%	15.5%	18.7%	23%	29.8%

Résultats qualitatifs (MD/FRMG 2008)

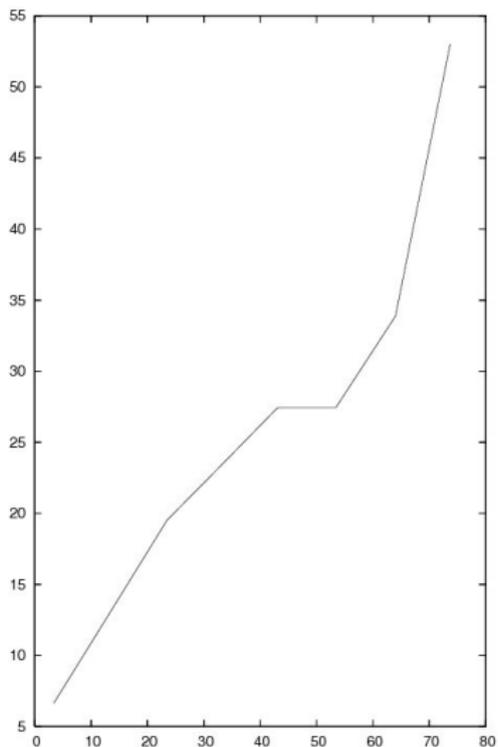
Rang	Token(s)/forme	$S_f^{(200)}$	$ \mathcal{O}_f $	err(f)	M_f	cause (de 2005)
1	:	78%	59992	98%	8.62	pb de segmentation sur : dans SxP
2	*	92%	1799	100%	6.90	corpus: phrases comportant le seul
3	Brève/brève	86%	554	100%	5.43	
4	Cf/cf	79%	858	100%	5.33	
5	etc/etc.	62%	365	100%	3.68	donner qqch de raisonnable à etc da
6	op/hop	63%	262	100%	3.51	
7	(...)/(...)	46%	1364	83%	3.29	
8	rectificatif	69%	114	98%	3.27	
9	Ibid	69%	91	100%	3.11	
12	65%	59	100%	2.64	ajout de comme ponctuation fina
13	XXe	46%	295	92%	2.63	
14	tentés	65%	57	100%	2.62	
15	apparue	64%	58	100%	2.61	
16	Nombre/nombre	56%	99	91%	2.59	
17	terminée	61%	69	96%	2.56	
18	apparu	58%	82	100%	2.56	
19	peut-on/Peuton	52%	129	95%	2.55	vérifier ce que fait SxPIPE sur peut-
20	intervenu	67%	45	100%	2.54	
100	Chiffres	41%	41	95%	1.54	
101	échéant	39%	51	82%	1.53	
102	Peut-être/peut	32%	121	97%	1.52	
103	Peut-être/être	32%	121	97%	1.52	
104	vienne	42%	34	85%	1.49	
105	sous-commandant/sous-__	37%	56	96%	1.49	
106	amÉrique/Amérique	43%	31	45%	1.48	
107	Rapport	29%	170	84%	1.48	
108	Andes/Ande	48%	22	100%	1.48	manque Andes comme np pluriel da

- 1 Contexte
- 2 Premières améliorations
- 3 Fouille d'erreurs
- 4 Suggérer des corrections d'erreurs**
- 5 Fouille d'erreurs et apprentissage (non supervisé)

- Remplacer un suspect par un **joker** (où mot inconnu) dans les phrases échec associées
- Ré-analyse avec **FRMG**
- Identifier les constructions syntaxiques récurrentes utilisées pour ancrer le joker dans les phrases
- Suggérer les informations lexicales pour le suspect

Note: ce programme est possible car **FRMG** peut retourner une forêt partagée de l'ensemble des analyses.

le taux de succès après ré-analyse avec joker corrélé avec le taux de suspicion.



Analyzing correction suggestions

2045 Seule/seule 1/18/v
 26 fiche/fiche 2/18/adj
 1612 Seuls/_Uw 16/18/adj
 307 confiance/confiance 16/18/adj
 2377 demain/demain 2/15/adj
 5245 Beaucoup/beaucoup 0/15/v
 5481 Certaines/_Uw 15/15/adj
 284 date/date 18/15/adj
 423 autorisée/autorisée 13/15/v
 5244 Beaucoup/_Uw 0/15/v
 2351 révisée/révisée 1/14/v
 5368 Pas/pas 2/14/v
 340 jour/jour 16/14/adj
 246 prospères/prospères 0/14/adj
 5358 Pas/pas 2/14/v
 4165 existe/existe 3/14/v
 316 etc/etc 0/14/v
 6 Desclée/_Uw 1/14/adj
 3423 Quoi/quoi 10/13/v
 5106 Amérique/Amérique 3/13/v
 1078 sentir/sentir 12/13/v
 3006 atout/atout 4/13/v
 481 taliban/taliban 1/13/adj
 3470 Êtant/étant 1/13/v
 795 censé/censé 0/13/v
 3422 Quoi/_Uw 10/13/v
 3468 Êtant/_Uw 1/13/v
 311 dus/dus 0/12/adj
 2394 Sinon/_Uw 0/12/adj

Enter if (or :rank) 246 id=246 rank=29

[edit comment](#)

[eric] identification correcte adjectif

info on 246: **prospères** /prospères

[_] Key/Lex => prospères/prospères Original Results => 0 success, 19 failures, 0 timeouts Best new results => 14

- [-] **_error_adj** Status: DONE Results: 14 success, 5 failures, 0 timeouts Type: detected
 - [+] Sentences:
 - [-] Hypothesis:
 - 27) Cat: adj Points: 2.24333833330257
 Relations: v => suspect (comp)
 - 57) Cat: adj Points: 1.98496851143454 Graph Dep: 1311 1653
 Relations: nc => suspect (N) suspect => prep (adjP)
 - 3) Cat: adj Points: 1.08043360070043 Graph Dep: 57 87 27
 Relations: nc => suspect (N)
 - 87) Cat: adj Points: 0.893071869162678 Graph Dep: 1653
 Relations: nc => suspect (N) suspect => adj (adj)
 - 1653) Cat: adj Points: 0.390905784728938
 Relations: nc => suspect (N) suspect => prep (adjP) suspect => adj (adj)
 - 47027) Cat: adj Points: 0.146689166651287
 Relations: v => suspect (object) suspect => det (det) suspect => adj (N)
 - 1311) Cat: adj Points: 0.040695516245616
 - 11) Cat: adj Points: 0.0366672916628218
 Relations: adj => suspect (N)
- [+] **_error_nc** Status: DONE Results: 7 success, 12 failures, 0 timeouts Type: detected

- 1 Contexte
- 2 Premières améliorations
- 3 Fouille d'erreurs
- 4 Suggérer des corrections d'erreurs
- 5 Fouille d'erreurs et apprentissage (non supervisé)**

Étant donné des paradigmes morphologiques:

- 1 choisir en corpus un large ensemble de formes
- 2 rechercher les lemmes potentiels pour chaque forme
⇒ lemmes en compétition
- 3 ordonner localement les lemmes sur la base d'un ratio, en démarrant avec une distribution initiale pour les lemmes
- 4 calcul d'une nouvelle distribution pour les lemmes
- 5 répéter à partir de l'étape 3 jusqu'à convergence

ACI "Masses de données": Traitement de descriptions botaniques (flores)



Corpus "Flore du Cameroun" (1963 – 2001)

Volumes	Pages	Av. Pages	Mots	Taxons
31	9466	305	1.5M	~ 2400

Tâches:

- Préparation des corpus: correction orthographique (OCR), structuration logique
- Traitements linguistiques préliminaires: morpho-syntaxe
- Extraction terminologique & expérience relations "gouverneur-gouverné"
- Extraction "Ontologie": analyse \Rightarrow dépendances syntaxiques + hypothèse **Harris**: des contextes syntaxiques similaire suggèrent des similarités sémantiques
 \Rightarrow **lancéolé** (adj) : **forme de feuille**
- Fouille de texte: obtenir les propriétés de chaque taxon
parsing + désambiguïsation grâce à l'ontologie

Tige simple ou ramifiée, robuste, feuillée.
Feuilles distiques, condupliquées, groupées en éventail, fortement coriaces, inégalement bilobées.
Inflorescence axillaire, en racème, en corymbe, paniculée, plus courte que les feuilles.
Fleurs non résupinées, petites, charnues.
Sépales et pétales dissemblables.
Labelle charnu, épais, sacciforme ou avec un éperon court, immobile.
Gynostème court, massif, avec deux stéolidies digitiformes, peu développées.
Stigmate ovale, profondément concave.
Anthère incombante, operculée, à parois minces.
Pollinies 4, unies en 2 paires, globuleuses, de taille inégale.
Rostelle court, digitiforme, charnu, obtus. Viscidie et tegula uniques.
Viscidie petite, ovale, assez épaisse.
Tegula linéaire, mince.
Restes du rostelle tronqués, foveolés au sommet.
– PL 282, p. 669.
Genre comprenant environ 10 espèces largement distribuées en Asie du Sud-Est, en Afrique tropicale et subtropicale, à Madagascar et dans les îles de la partie occidentale de l'Océan Indien.

Besoin

- d'améliorer la grammaire et le lexique pour un domaine spécialisé
fouille d'erreur & 14rounds: augmentation de la couverture de 36% à 67%
- Désambiguïsation des catégories syntaxiques:
eg.: **trident**: *noun* en général, *adj* dans Biotim
- Désambiguïsation des dépendances
pour appliquer l'hypothèse distributionnelle

- 1 pour chaque forme, démarre avec des poids initiaux pour les catégories syntaxiques et pour les dépendances entrantes:
 - ▶ 1 si non ambiguë
 - ▶ ratio si ambiguë
- 2 calcul poids moyen sur l'ensemble du corpus
- 3 répétition des étapes 1 et 2 avec la nouvelle distribution jusqu'à convergence

- 1 pour chaque forme, démarre avec des poids initiaux pour les catégories syntaxiques et pour les dépendances entrantes:
 - ▶ 1 si non ambiguë
 - ▶ ratio si ambiguë
- 2 calcul poids moyen sur l'ensemble du corpus
- 3 répétition des étapes 1 et 2 avec la nouvelle distribution jusqu'à convergence

L'algorithme est clairement une approximation grossière:
Ajuster les poids sur une forme devrait prendre en compte le contexte d'une analyse.

Using:

- Désambiguisation
- Hypothèse distributionnelle
- Marqueurs linguistiques
- Quelques termes germes

organes	propriétés	autres
nervure	oblong	diamètre
fleur	ovale	longueur
face	ovoïde	hauteur
feuille	elliptique	largeur
limbe	glabre	taille
rameau	lancéolé	forme
sommet	ellipsoïde	forêt
sépale	globuleux	* d
foliole	floral	couleur
base	aigu	* mètre

- 100 premiers “organes” OK, sauf **dessus**, **dessous**, **bord coté** et quelques pronoms comme **une**
- 100 premières “propriétés” parfaites sauf peut-être **semblable**
- 100 premiers “autres” suggèrent des marqueurs linguistiques, mais incluent quelques “organes” comme **rhizome**

Obtention de graphes conceptuels

Menu ToolBar
Archivo Editar Ayuda

Consulta: inflorescence courte

Resultado

Resultado de la consulta

- Acampe
- Acampe.E1
- Acampe.E2
- Acampe.E3
- Acampe.E4
- Acampe.E5: **Inflorescence axillaire , en racème , en corymbe , paniculée , plus courte que les feuilles**
- Acampe.E6
- Acampe.E7
- Acampe.E8
- Acampe.E9
- Acampe.E10
- Acampe.E11
- Acampe.E12
- Acampe.E13
- Acampe.E14
- Acampe.E15
- Acampe.E16

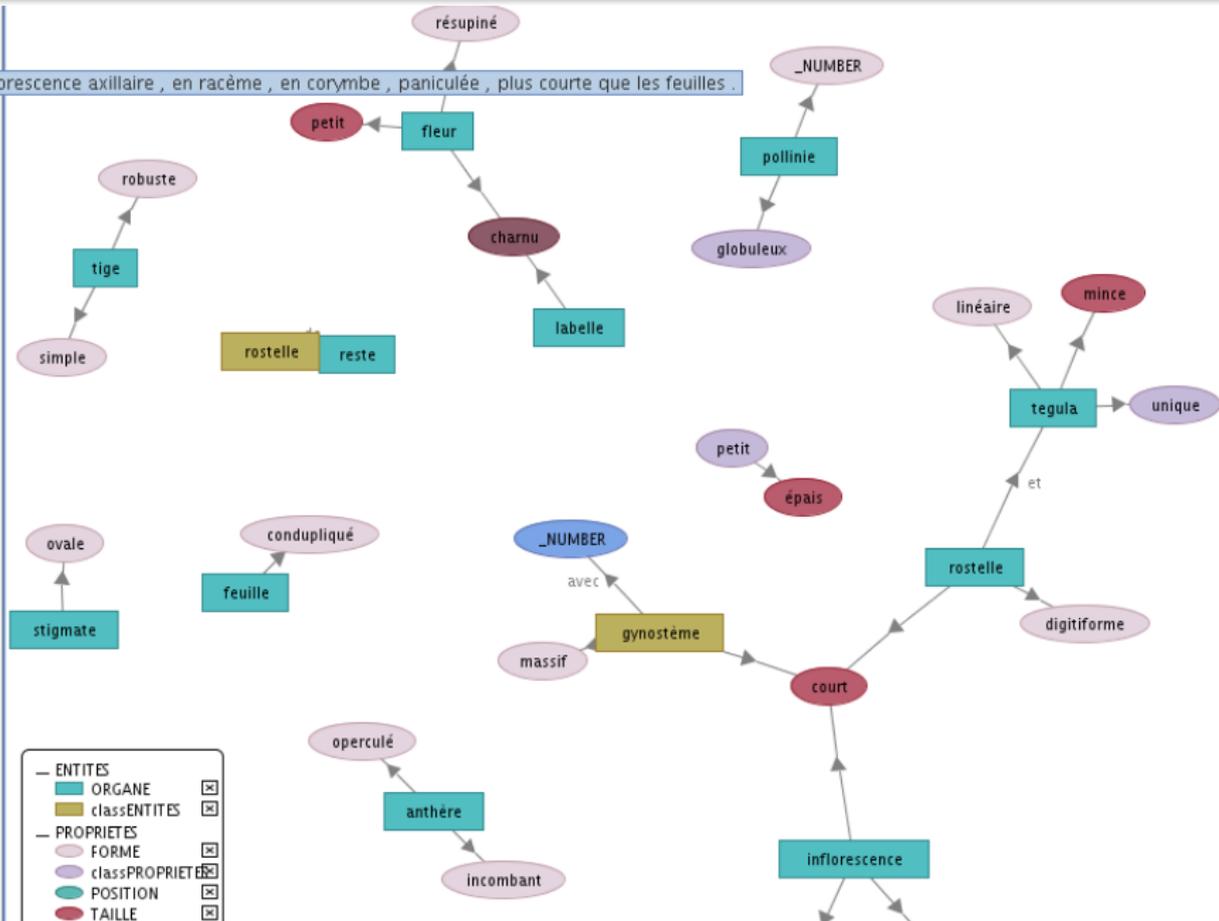
- Ancistrorhynchus_serratus
- Ancistrorhynchus_schimus
- Anthoanthes
- Bellis_biedia_acuta
- Canna_bis
- Cyphothemma_suakense
- Cyrtorchis
- Dicellandra
- Disperis_nitida
- Egglegia
- Kryptanthes_tetrazeyana
- Mamocyan_calophyllum
- Mossaea_katesii
- Nephralingis
- Ptilostigma_reticulatum
- Platodiscus
- Strychnos_jinnocua
- Strychnos
- Tetelia_alfelli
- Toddalopsis_heterophylla
- Vegetis_lavilii

— ENTITES
— PROPRIÉTÉS

- ENTITES: organe (rectangle), classe ENTITES (carré), forme (carré), position (carré), taille (carré), texture (carré), autre (carré)
- PROPRIÉTÉS: forme (carré), classe PROPRIÉTÉS (carré), position (carré), taille (carré), texture (carré), autre (carré)

Obtention de graphes conceptuels

pe.E3: inflorescence axillaire, en racème, en corymbe, paniculée, plus courte que les feuilles.



- ENTITES
 - ORGANE
 - classENTITES
- PROPRIETES
 - FORME
 - classPROPRIETES
 - POSITION
 - TAILLE

données observées	données cachées
phrases succès/échec	forme erronée
phrases	bonne cat. syntaxique
phrases	bonne séquence tag
phrases	bon arbre d'analyse

Problème: retrouver les bonnes données cachées expliquant les données observées.

La fouille d'erreurs (*Error mining*) est un cas particulier de **Expectation Maximization** ?

[Dempster, Laird & Rubin]

Étant donnée des **données observées** \mathcal{X} avec des **données cachées** \mathcal{Y} , et supposant une distribution paramétrique $p(x, y|\theta)$, rechercher θ maximisant **log-likelihood**:

$$L(\theta|\mathcal{X}, \mathcal{Y}) = \prod_{i=1}^{i=N} p(x, y|\theta)$$
$$\theta^* = \underset{\theta}{\operatorname{argmax}} \log L(\theta|\mathcal{X}, \mathcal{Y})$$

Calcul itératif, améliorant le *log-likelihood* pour $\theta^{(i)}$ à chaque étape

- **Expectation** – niveau local

$$Q(\theta, \theta^{(i)}) = E[\log p(\mathcal{X}, \mathcal{Y}|\theta) | \mathcal{X}, \theta^{(i)}]$$
$$= \sum_y \log \left(\prod_x p(x, y|\theta) f(y|x, \theta^{(i)}) \right)$$

- **Maximization** – niveau global

$$\theta^{(i)} = \operatorname{argmax} Q(\theta, \theta^{(i)})$$

- **Baum-Welsh** variant de **Forward-Backward** pour les **Modèles Cachés de Markov (HMM)**
 - ▶ Amélioration d'étiqueteurs
 - ▶ Fouille d'erreurs reformulé en HMM **Yvon**
 - ★ chaque phrase échec change d'état de 'S' vers 'F' au passage du suspect
 - ★ trouver le point le plus probable de changement
 - ★ ⇒ suggère certaines extensions
- **Désambiguisation d'analyses syntaxiques**
 - ▶ Approximation **Viterbi**: estimation sur la meilleure analyse
 - ▶ Algo. de Programmation Dynamique sur les forêts partagées pour calculer l'expectation
 - ▶ mais semble ne pas très bien marcher (sensibilité aux conditions initiales, optimum local)

- modèle statistique simple + interface \Rightarrow bonne efficacité et interactivité
- applicable à d'autres sortes d'information, jeu entre les niveaux locaux et globaux
- meilleur compréhension du modèle et de possibles généralisations \Rightarrow Expectation Maximization
- premiers pas vers du *bootstrap* des outils et des ressources

- développement au long terme de ressources
- exploitation des ressources sur corpus avec des outils
- exploitation du retour (*feedback*)
- exploration et validation avec des interfaces intégrées
- audience élargie grâce à des interfaces WEB collaboratives telles que **EASYREF** pour le TreeBank Easy