PASSAGE

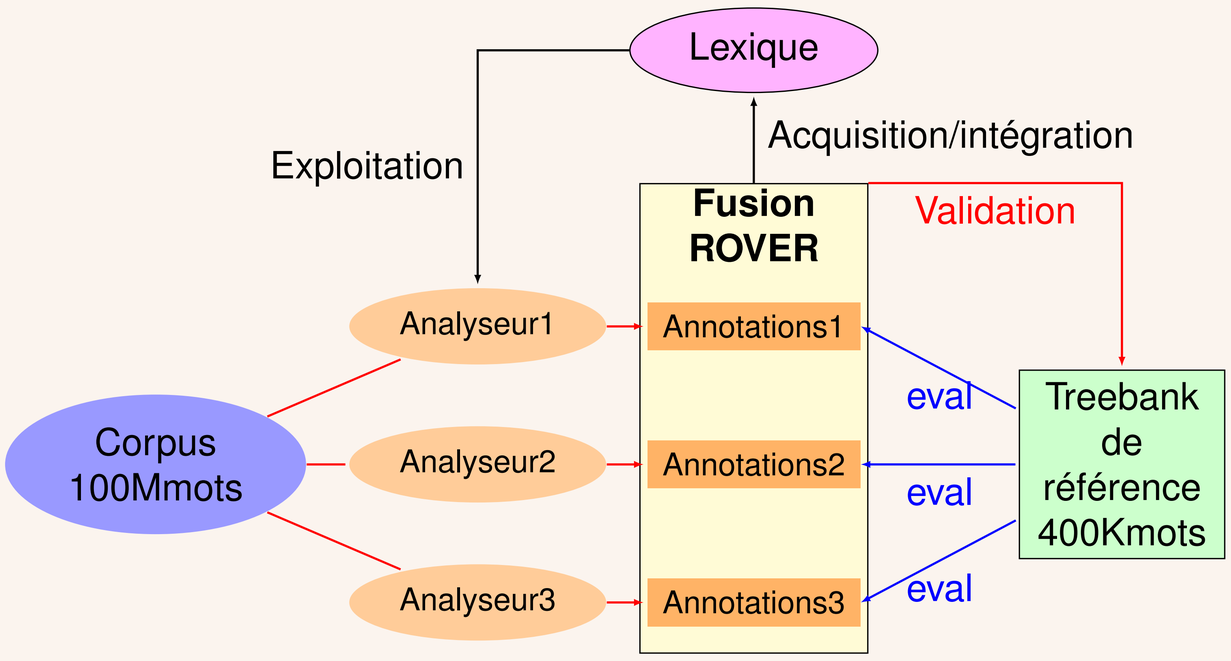
Produire des annotations syntaxiques à grande échelle

http://atoll.inria.fr/passage

Éric Villemonte de la Clergerie
Eric.De_La_Clergerie@inria.fr
INRIA Paris-Rocquencourt

ANR 2006 MDCA-013
Grand Colloque STIC 2007
Paris, 5-7 Novembre 2007

Établir un cercle vertueux entre outils et ressources linguistiques



Exploiter de très grands corpus

Les corpus annotés syntaxiquement comme le **Penn Tree Bank** sont très importants pour le Traitement Automatique des Langues [**TAL**] mais **rares** et **coûteux** à développer.

Par contre, il est maintenant possible d'accéder à de très grandes quantités de textes électroniques en français:

Corpus	Taille	Nature
Corpus EASy	1Mmots	multi-styles
Wikipedia Fr	$\sim$ 86Mmots	encyclopédique collaboratif libre
Wikisources	$\sim$ 80Mmots	littéraire libre
Monde Diplomatique	18Mmots	journalistique
FRANTEXT	20Mmots	littéraire libre
Europarl	28Mmots	débat Parlement européen
JRC-Acquis	39Mmots	juridique européen
Corpus Ester	1Mmots	oral transcrit
Total (actuel)	$>$ **270 Mmots**	

Profiter d'une dizaine d'analyseurs

Une occasion unique, source de diversité (formalismes, technologies, ...)

Analyseur	Origine	Nature
FRMG	INRIA	TIG/TAG+DyALog
SxLFG	INRIA	LFG+Syntax
LLP2	LORIA	TAG
LIMA	CEA-LIST	
TagParser	TAGMATICA	Induction + règles
Syntex	ERSS	Système de règles
GP1 & GP2	LPL	Grammaires de Propriétés
Cordial	SYNAPSE	
SYGMART	LIRMM	
XIP	XRCE	Cascade de règles

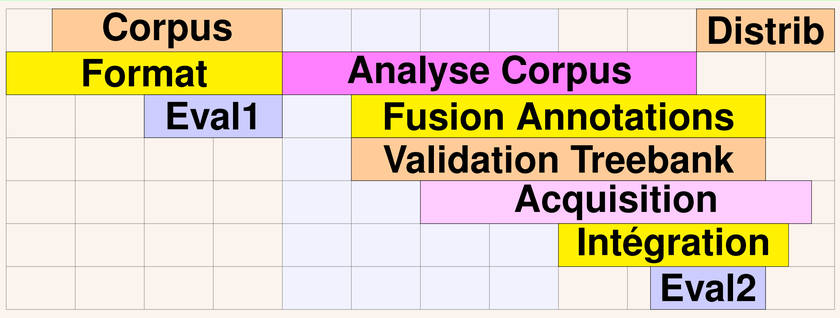
Évaluer et combiner les annotations

- Expertise de Technolangue **EVALDA/EASy** sur l'évaluation des analyseurs du français:
  - **[Eval1]** Campagne Novembre 2007
  - **[Eval2]** Campagne fin 2009
- Format EASy: chunks + dépendances
  - Ajout des groupes récursifs (constituance)
  - Vers le respect des **standards** ISO TC37SC4.
- **ROVER**: Fusion des jeux d'annotations paramétrée par les évaluations + *feedback*
- Validation manuelle d'un sous-corpus
- $\Rightarrow$ Améliorer les analyseurs !

Acquérir et intégrer des connaissances

- Exploiter les annotations ROVER pour de l'acquisition de **connaissances lexicales**:
  - information de valence (verbes, ...)
  - classes d'alternation verbales (Levin)
  - probabilités de désambiguation catégories, constructions, ...
  - restrictions de sélection
  - classes sémantiques (Hyp. distributionnelle Harris)
  - morphologie dérivationnelle avec transfert des informations syntaxiques
- Les valider et intégrer dans un lexique
- Les exploiter dans certains analyseurs $\Rightarrow$ Améliorer les analyseurs !

Calendrier indicatif



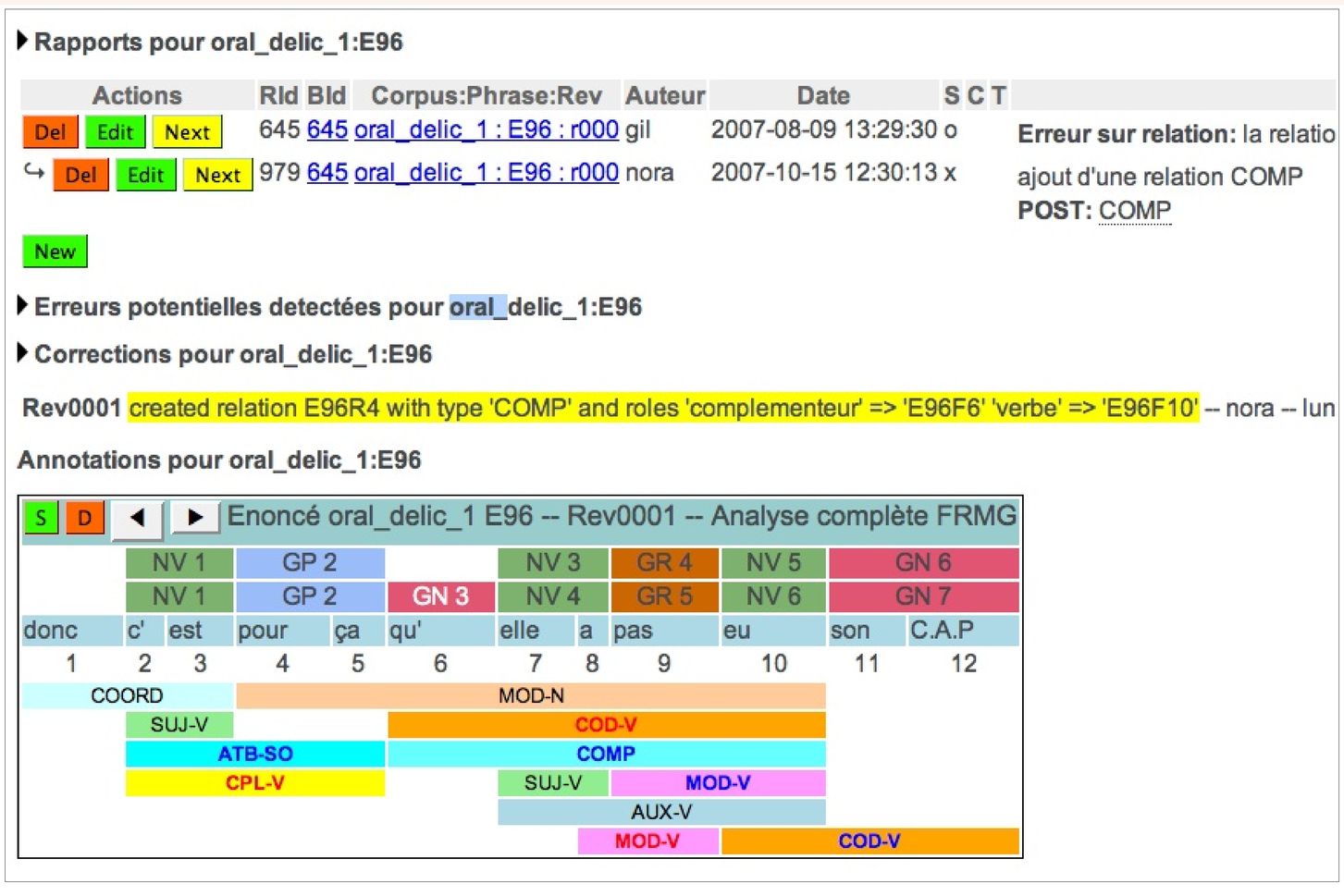
Retombées attendues

- Rendre librement disponible des ressources linguistiques de qualité pour le français:
  - des corpus annotés, dont
    - le **ROVER** ($>$ 100 Mmots)
    - un ***treebank*** vérifié manuellement (400Kmots)
  - des ressources lexicales
- Améliorer:
  - les analyseurs syntaxiques robustes efficaces du français
  - les méthodologies d'évaluation et de fusion des annotations
  - les méthodologies d'acquisition par bootstrap
- Permettre des applications par analyse syntaxique de corpus:
  - Acquisition de connaissances (extraction d'ontologies, ...)
  - Extraction d'informations (fouille de textes, veille, question-réponse, ...)

Infrastructure pour passer à l'échelle

Besoin d'une infrastructure solide pour gérer des masses de données

- Utilisation de *fermes* de machines pour analyser de gros corpus
- EasyRef, un prototype WEB de gestion collaborative d'annotations visualisation, recherche, édition, *versionning*, comparaison, dépôt, fusion, ...



INRIA LIR/LIMSI LORIA CEA-LIST ELDA TAGMATICA XRCE LPL ERSS SYNAPSE LIRMM