ANR 2006 MDCA

PASSAGE

Produire des annotations syntaxiques à grande échelle

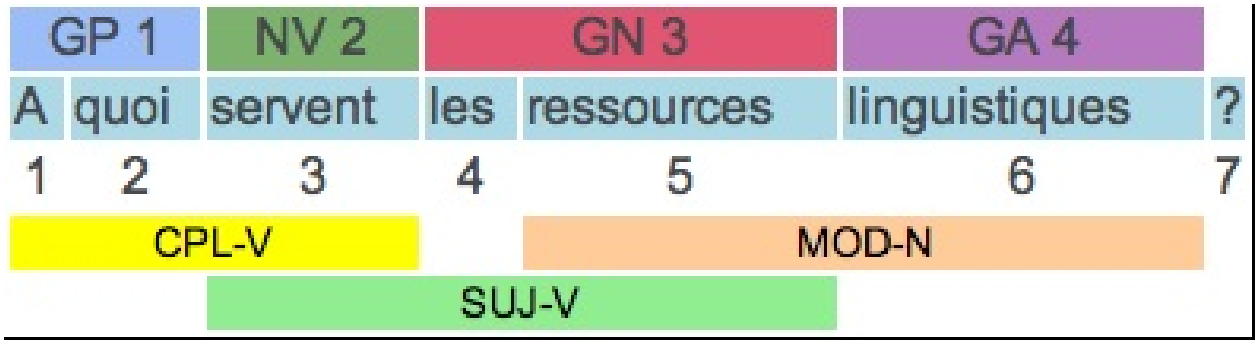
http://atoll.inria.fr/passage

Éric de la Clergerie
Eric.De_La_Clergerie@inria.fr

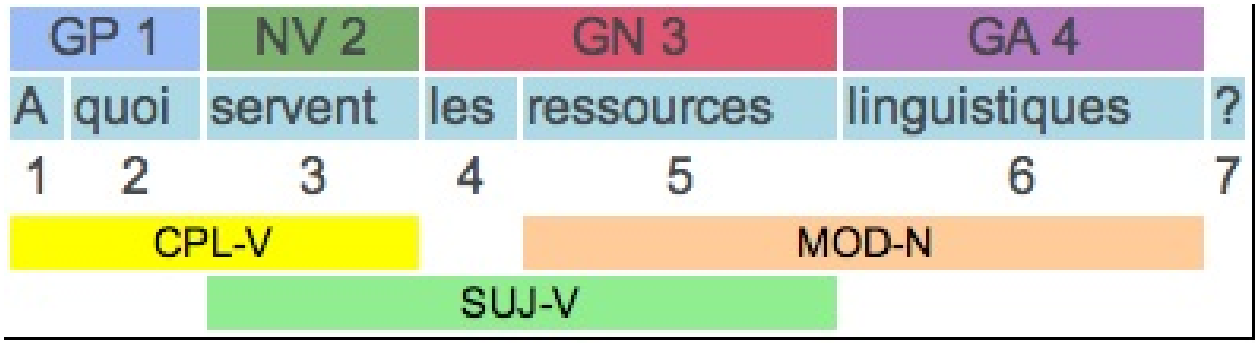
INRIA Paris-Rocquencourt

Grand Colloque STIC 2007
6 Novembre 2007

Améliorer le traitement linguistique du français



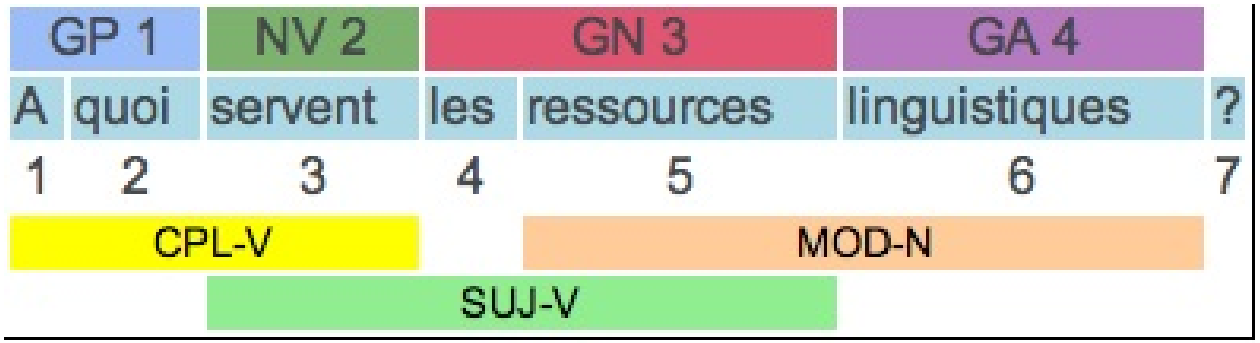
Améliorer le traitement linguistique du français



Traitement syntaxique du langage :

- pas de consensus sur les formalismes
- diversité des phénomènes syntaxiques
- les mots gouvernent la syntaxe
  $\Longrightarrow$ lexiques syntaxiques riches et couvrants : manque pour le français
- les corpus pour saisir les usages des mots
  $\Longrightarrow$ Tendance : corpus **manuellement** annotés syntaxiquement (TreeBank) pour amorcer le processus.

Améliorer le traitement linguistique du français



Traitement syntaxique du langage :

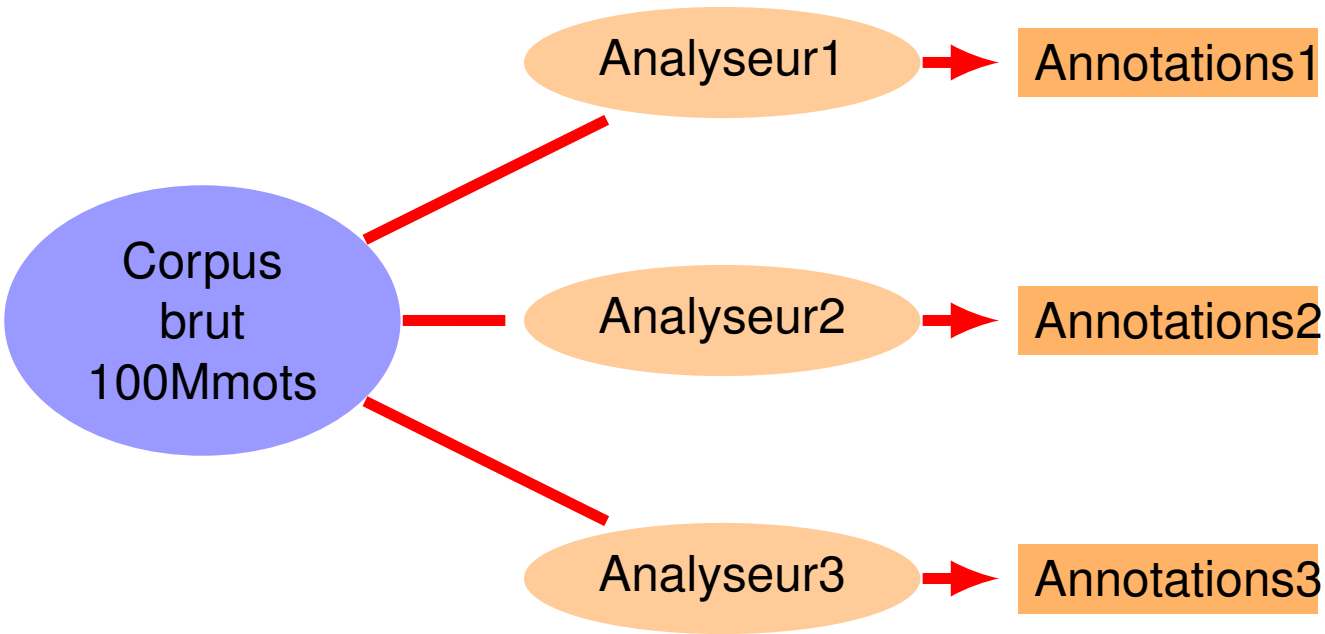
- pas de consensus sur les formalismes
- diversité des phénomènes syntaxiques
- les mots gouvernent la syntaxe
  $\Longrightarrow$ lexiques syntaxiques riches et couvrants : manque pour le français
- les corpus pour saisir les usages des mots
  $\Longrightarrow$ Tendance : corpus **manuellement** annotés syntaxiquement (TreeBank) pour amorcer le processus.

PASSAGE : proposition d'une approche plus intégrée

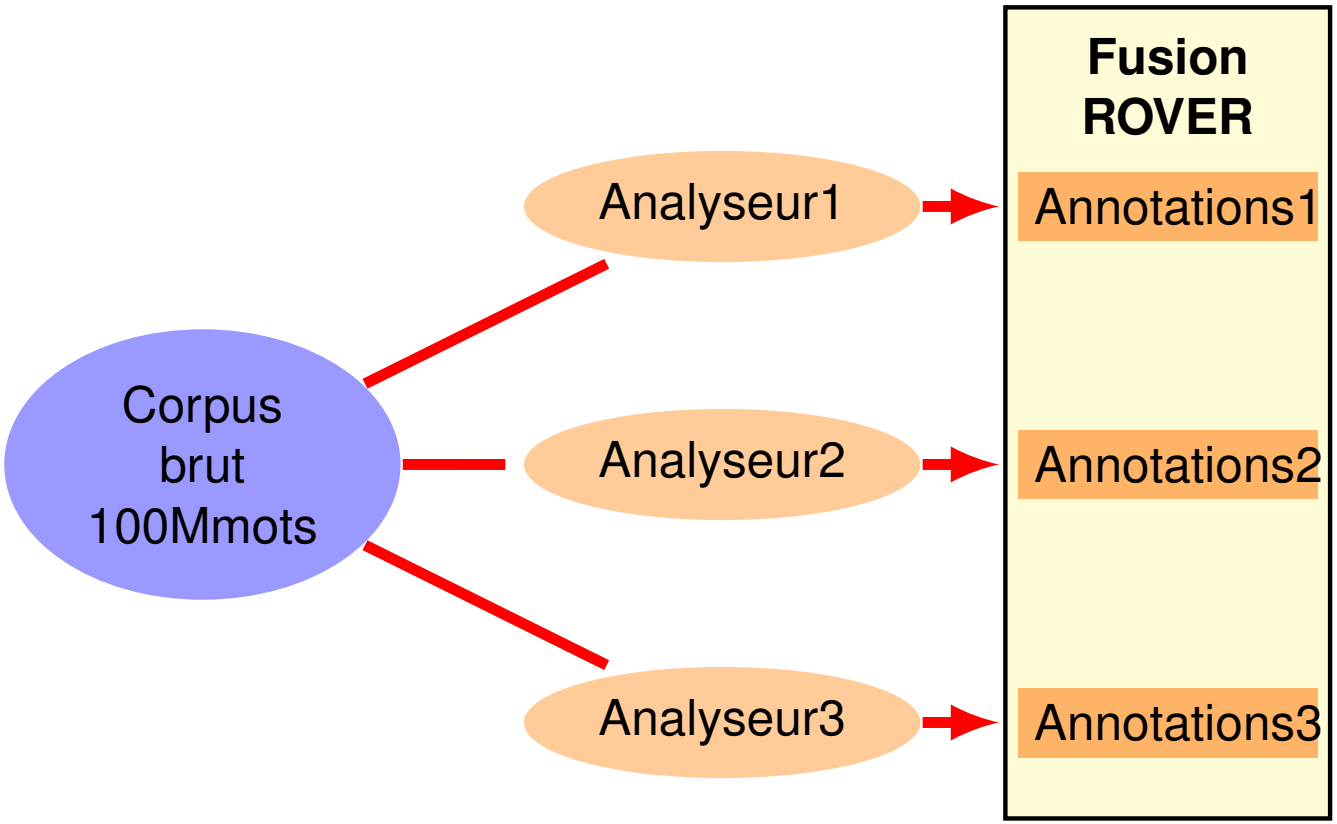
Établir un cercle vertueux entre outils et ressources

Corpus
brut
100Mmots

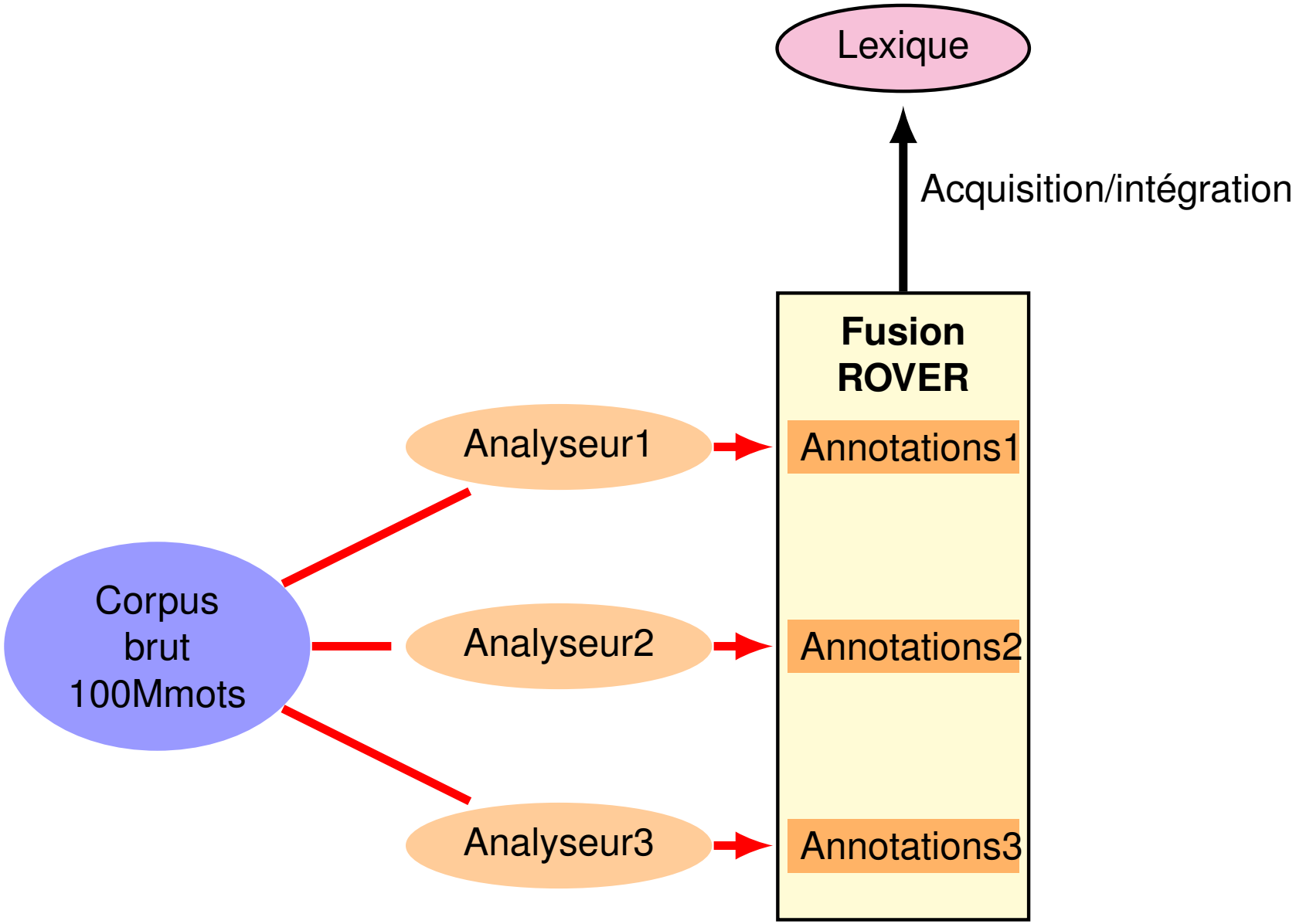
Établir un cercle vertueux entre outils et ressources



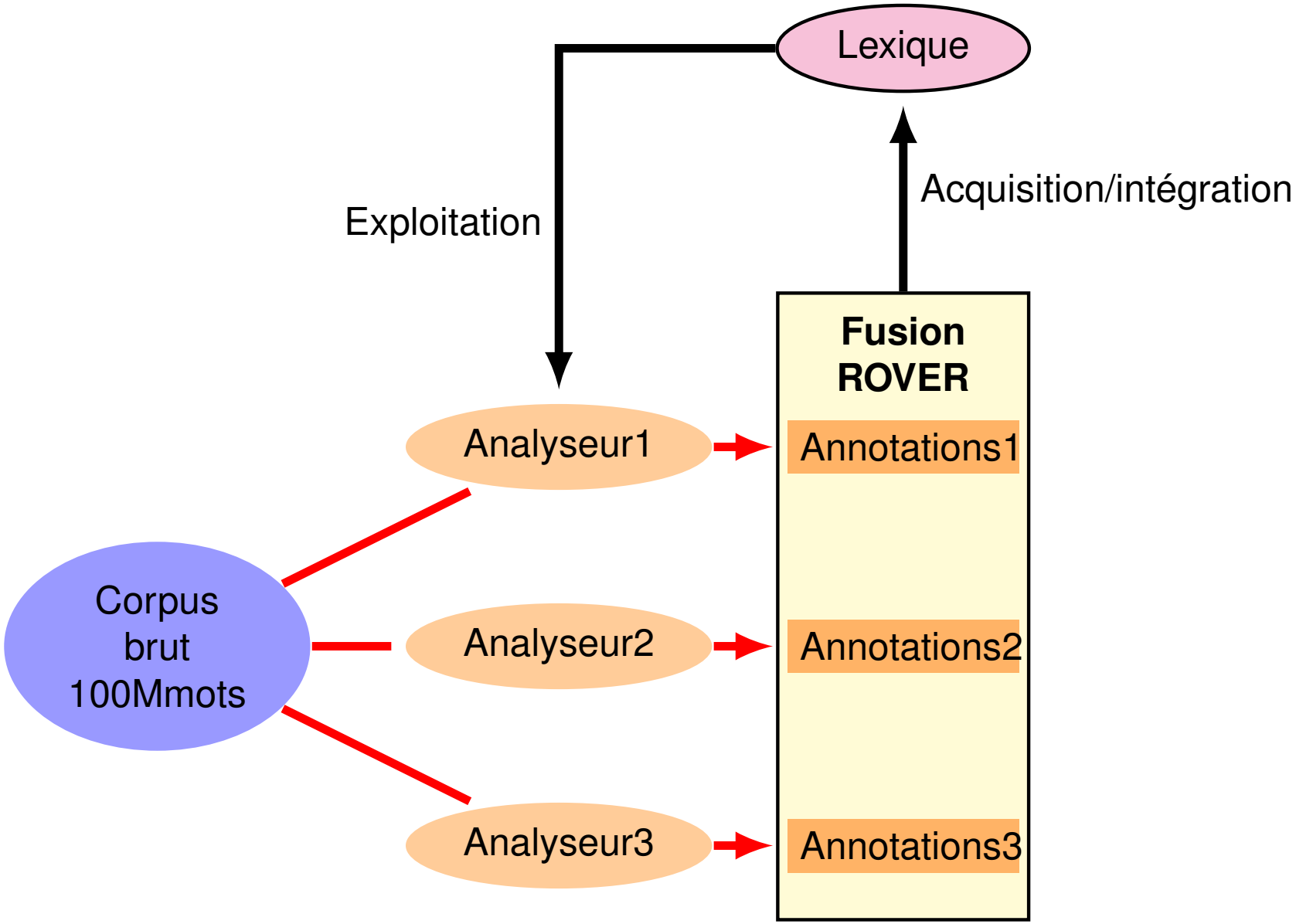
Établir un cercle vertueux entre outils et ressources



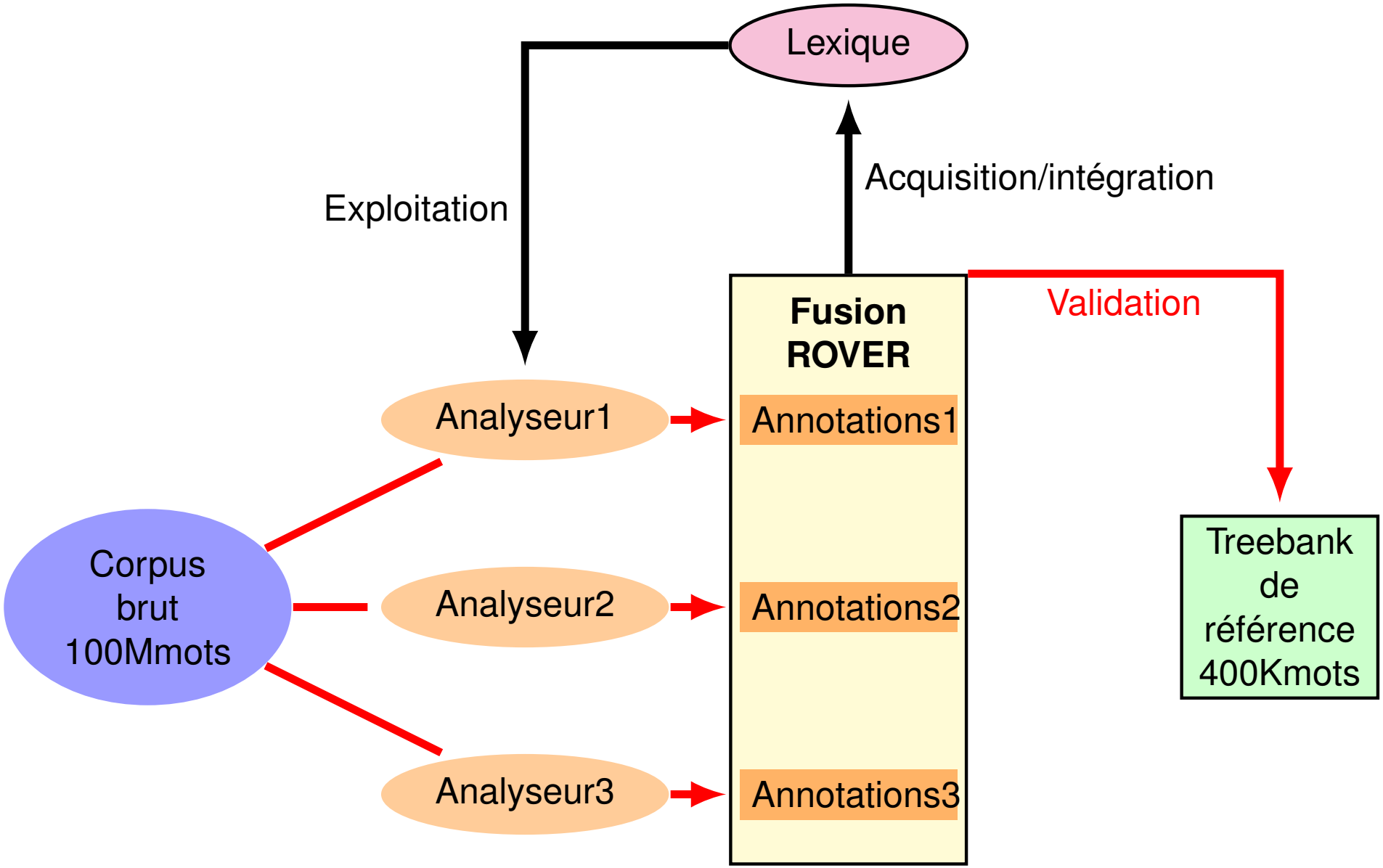
Établir un cercle vertueux entre outils et ressources



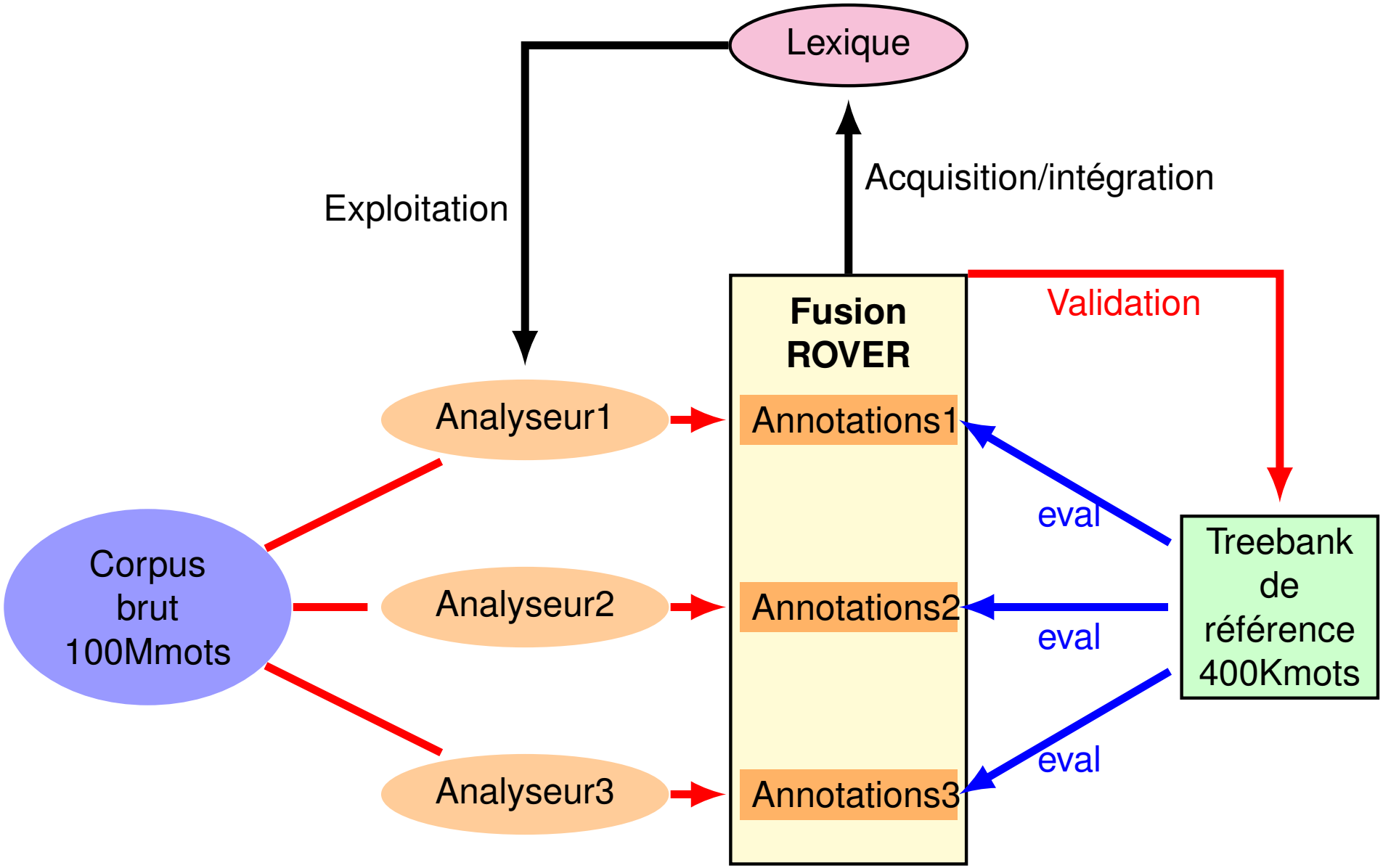
Établir un cercle vertueux entre outils et ressources



Établir un cercle vertueux entre outils et ressources



Établir un cercle vertueux entre outils et ressources



Consortium

ALPAGE
INRIA Paris7

TALARIS/LORIA

LIC2M/CEA-LIST

LIR/LIMSI

Consortium

ELDA

TAGMATICA

ERSS

ALPAGE
INRIA Paris7

LPL

TALARIS/LORIA

LIC2M/CEA-LIST

LIR/LIMSI

SYNAPSE

XRCE

LIRMM

Faire coopérer une dizaine de parseurs

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	Système de règles
TagParser	TAGMATICA	Induction + règles
Syntex	ERSS	Système de règles
GP1 & GP2	LPL	Grammaires de Propriétés
Cordial	SYNAPSE	
SYGMART	LIRMM	
XIP	XRCE	Cascade de règles

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.

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**	

Évaluer et combiner les annotations

- Expertise de Technolangue **EVALDA/EASy** sur l'évaluation des analyseurs du français :
  - **[Eval1]** Campagne Novembre 2007 (en cours)
  - **[Eval2]** Campagne fin 2009

- Format EASy :
  chunks (GN, GA, NV, ...) + dépendances (Sujet-Verbe, Objet-Verbe, ...)
  - Existence guide d'annotation assez complet
  - Existence $\sim$ 4K phrases annotées
  - Vers l'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

Acquérir et intégrer des connaissances

- Exploiter le ROVER pour acquérir des **connaissances lexicales** :
  - Information de valence (verbes, . . .)
    Sujet donne **Objet à-objet**
    Sujet le lui donne
  - Classes d'alternation verbales (Levin)
    Transfert de possession : vendre, céder, . . .
  - Probabilités de désambiguation
    catégories lexicales, constructions syntaxiques, restrictions de sélection, . . .
  - Classes sémantiques (Hyp. distributionnelle Harris)
    (humain $\vee$ organisation) annonce que Y
    humain est nommé à la tête de organisation
  - Morphologie dérivationnelle avec transfert des informations syntaxiques
    partir/départ ; X annoncer que Y /l'annonce de Y par X

- Les valider et intégrer dans un lexique

- Les exploiter dans certains analyseurs
  $\Longrightarrow$ Améliorer les analyseurs !

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 par exemple GRID 5000 ?
- EASYREF, un prototype WEB de gestion collaborative d'annotations visualisation, recherche, édition, *versionning*, comparaison, dépôt, fusion
  - Utilisé pour corrigé les annotations EASy existantes
  - Utilisé pour annoter 500 nouvelles phrases
  - Démo à STIC 2007

EasyRef

▸ **Rapports pour oral_delic_1:E96**

Actions	RId	BId	Corpus:Phrase:Rev	Auteur	Date	S	C	T	
Del Edit Next	645	645	oral_delic_1 : E96 : r000	gil	2007-08-09 13:29:30	o			**Erreur sur relation:** la relatio
↪ Del Edit Next	979	645	oral_delic_1 : E96 : r000	nora	2007-10-15 12:30:13	x			ajout d'une relation COMP **POST:** COMP

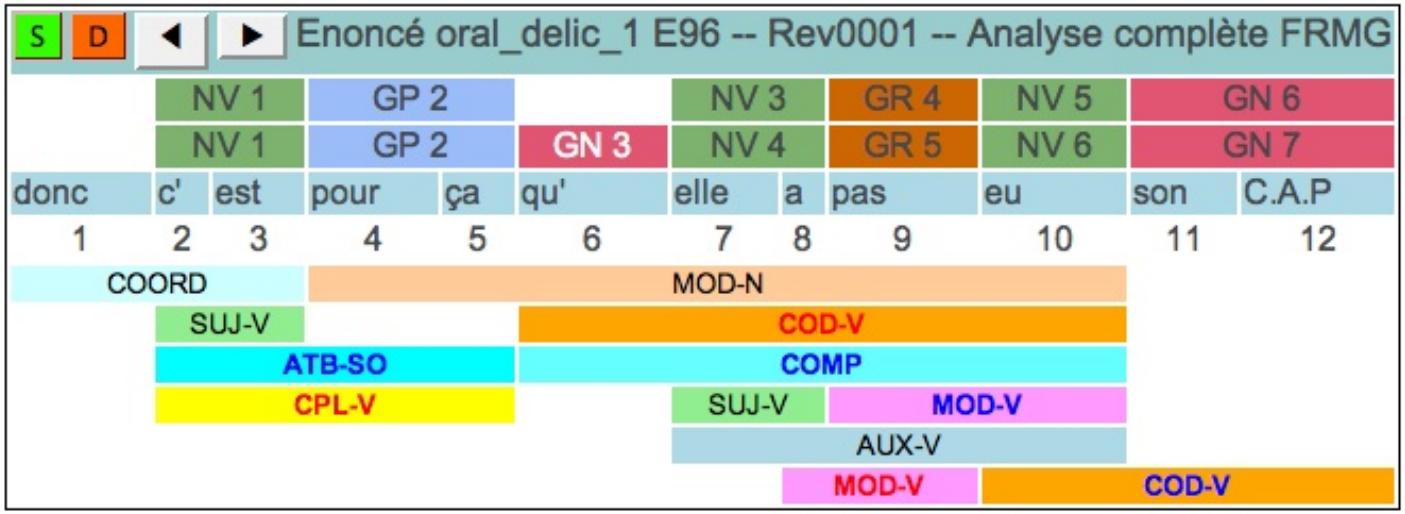
New

▸ **Erreurs potentielles detectées pour oral_delic_1:E96**

▸ **Corrections pour oral_delic_1:E96**

Rev0001 created relation E96R4 with type 'COMP' and roles 'complementeur' => 'E96F6' 'verbe' => 'E96F10' -- nora -- lun

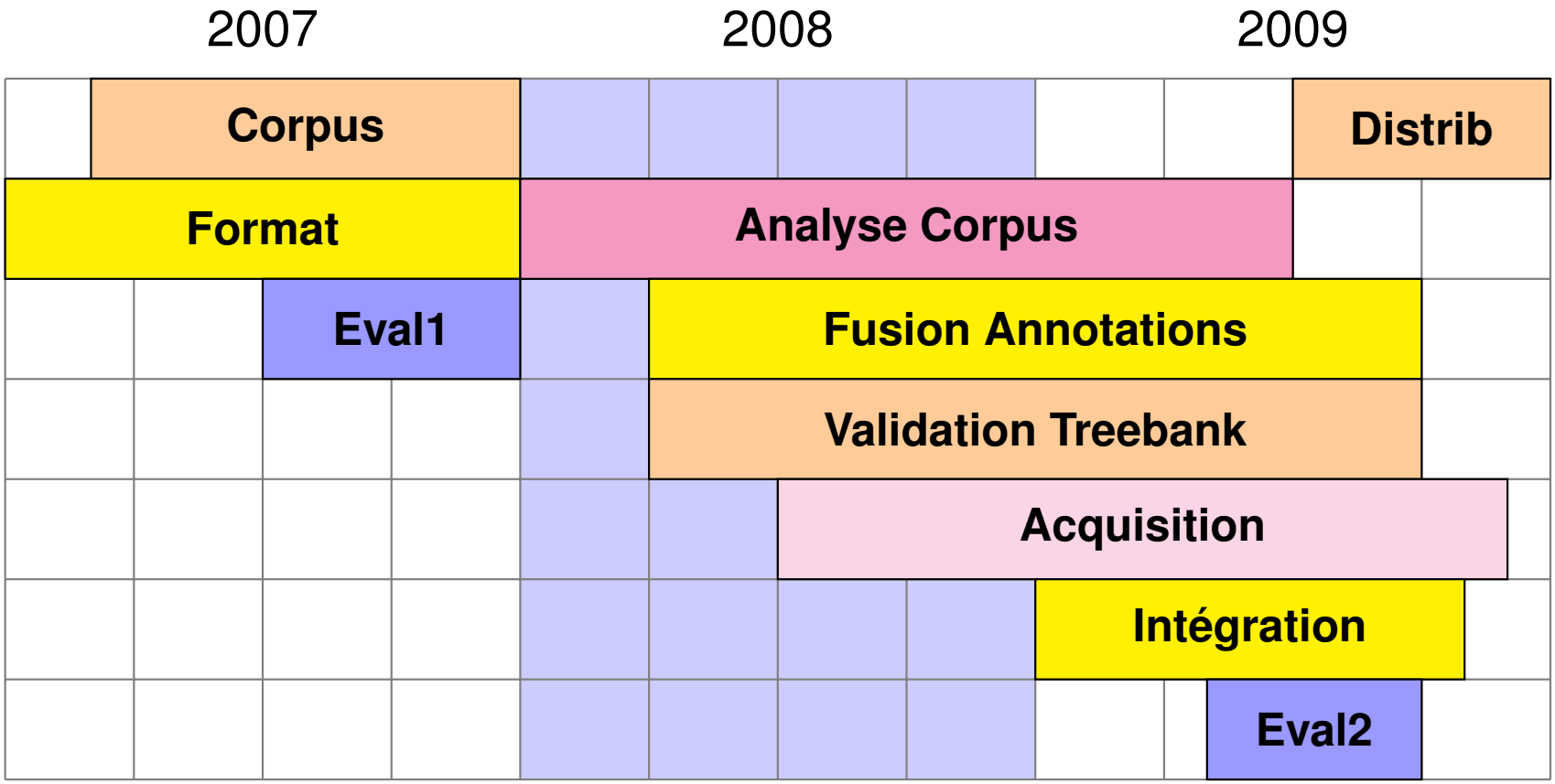
Annotations pour oral_delic_1:E96



Infrastructure : travailler ensemble

- Site Web : `http://atoll.inria.fr/passage`
- Liste de diffusion `passage@inria.fr`
- Zone wiki (DOKUWIKI)
  - Discussion et mise au point : corpus, protocoles, ...
  - Présentation et discussion : analyseurs
  - ...
- Zone INRIA GForge
  - Accès unifié aux logiciels et ressources
  - Versionning (SVN), gestionnaire de bugs, ...

Calendrier indicatif



Retombées attendues

- Rendre disponible des ressources linguistiques de qualité pour le français :
  - des corpus annotés, dont
    - ⋆ le **ROVER** ($>$ 100 Mmots)
    - ⋆ le ***TREEBANK*** vérifié manuellement (400Kmots)
  - des **ressources lexicales**

  Distribution : ELDA et CNRTL, en statut libre si possible

- 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, . . .)

PASSAGE pour aller de l'avant

PASSAGE pour permettre d'amorcer d'autres projets :

- Corpus de Référence du Français (**FNC**),
  dans la lignée de **BNC** et **ANC**
- Expériences de transfert syntaxique entre langues
  à partir de corpus alignés comme **Europarl**, **JRC-Acquis** et **Wikipedia**
  $\Longrightarrow$ import de l'anglais, export vers d'autres langues européennes
- Acquisition de connaissances, à partir de **Wikipedia**

Discussions sur PASSAGE $\Longrightarrow$ Co-organisation ICGL'08 Workshop
«*Automated Syntactic Annotations for Interoperable Language Resources*»
Hong-Kong, Janvier 2008