PASSAGE:

From French Parser Evaluation to Large Sized Treebanks

http://atoll.inria.fr/passage

Éric de la Clergerie (INRIA)
Djamel Mostefa (ELDA)
Patrick Paroubek (CNRS/LIMSI)
Olivier Hamon (ELDA-LIPN)
Christelle Ayache (ELDA)
Anne Vilnat (CNRS/LIMSI)

LREC'08
Marrakech, May 29th 2008

From EASy to Passage

EASy (2003–2006)
French Technolangue program
First French parsing evaluation campaign
15 parsers

From EASy to Passage

EASy (2003–2006)
French Technolangue program
First French parsing evaluation campaign
15 parsers

PASSAGE (2007–2009)
French ANR MDCA
Evaluation & much more
Dynamic Treebank

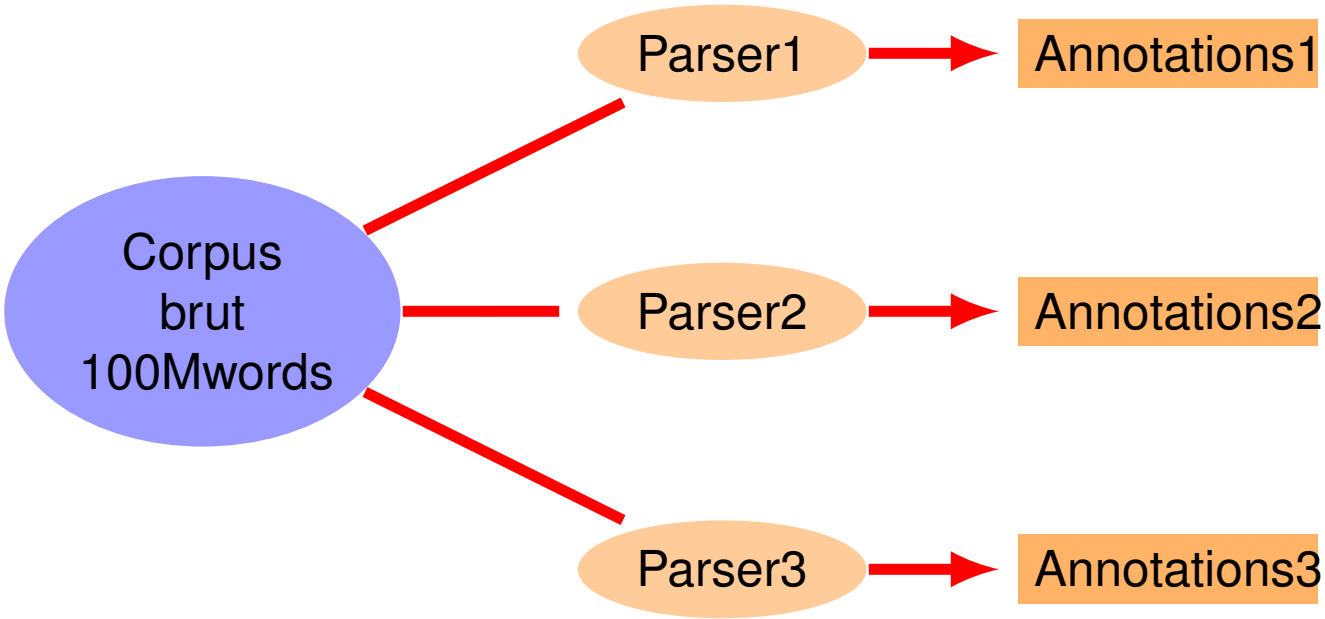
Benefits from **EASy** :

- Existence of several parsers for French
- These parsers are able to produce EASy annotations

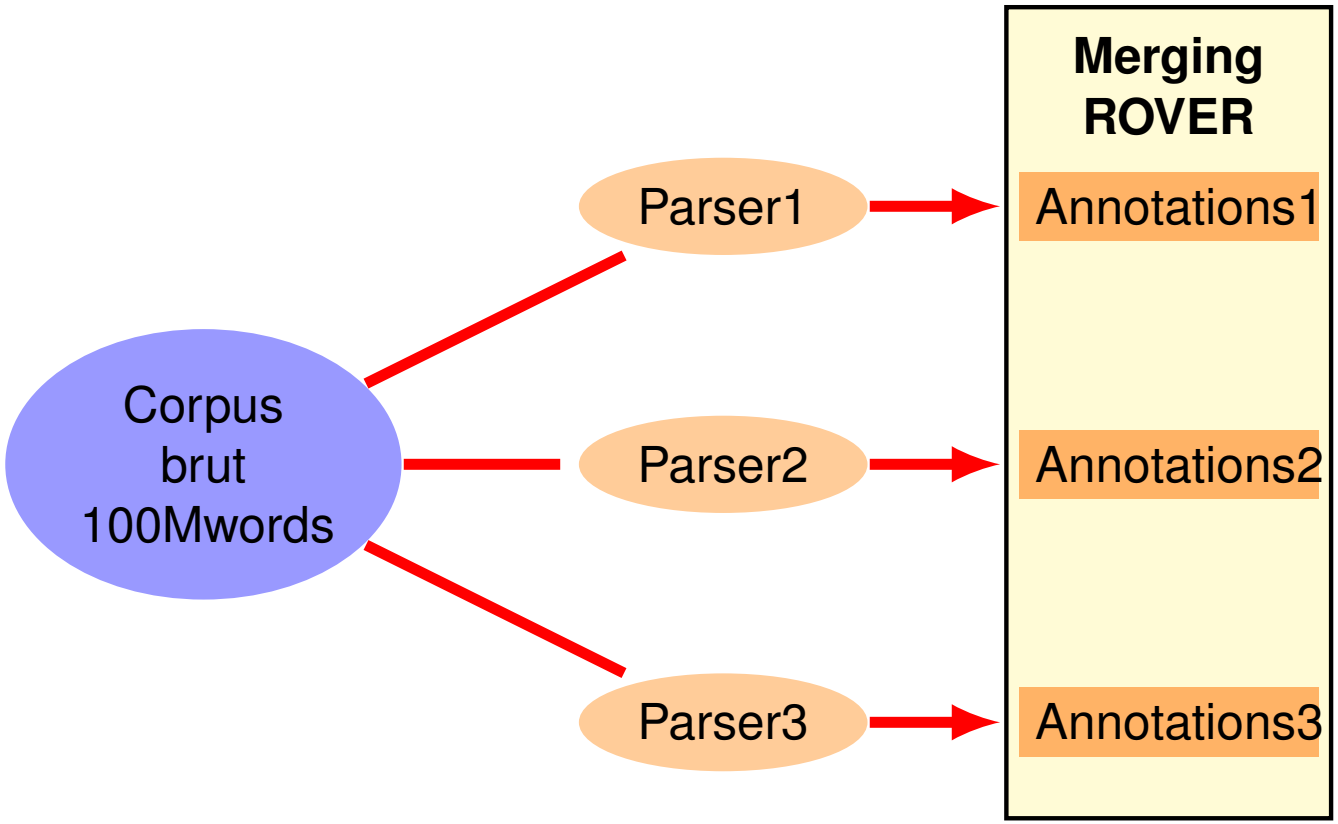
Entering a virtuous loop between tools and resources

Corpus
brut
100Mwords

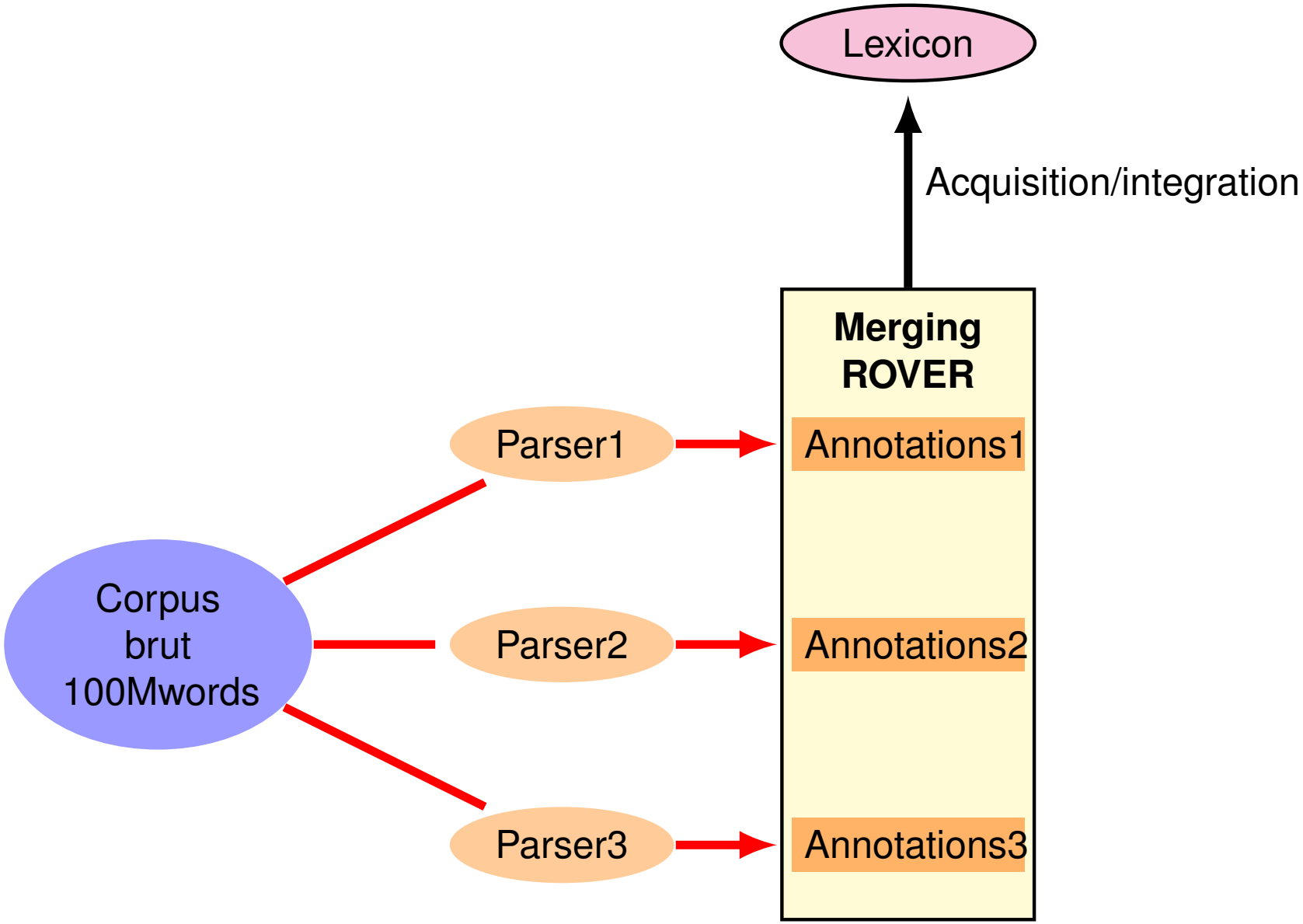
Entering a virtuous loop between tools and resources



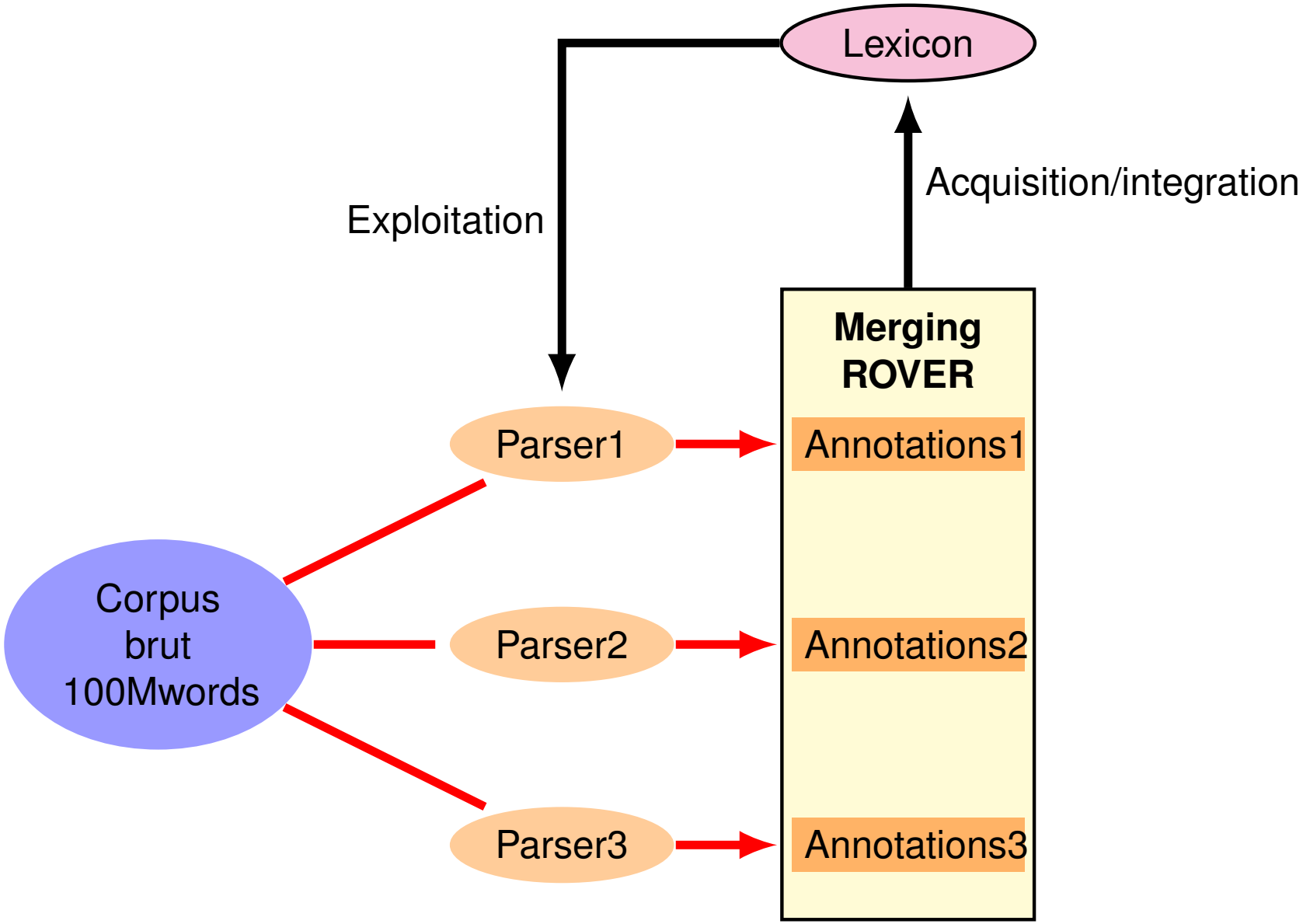
Entering a virtuous loop between tools and resources



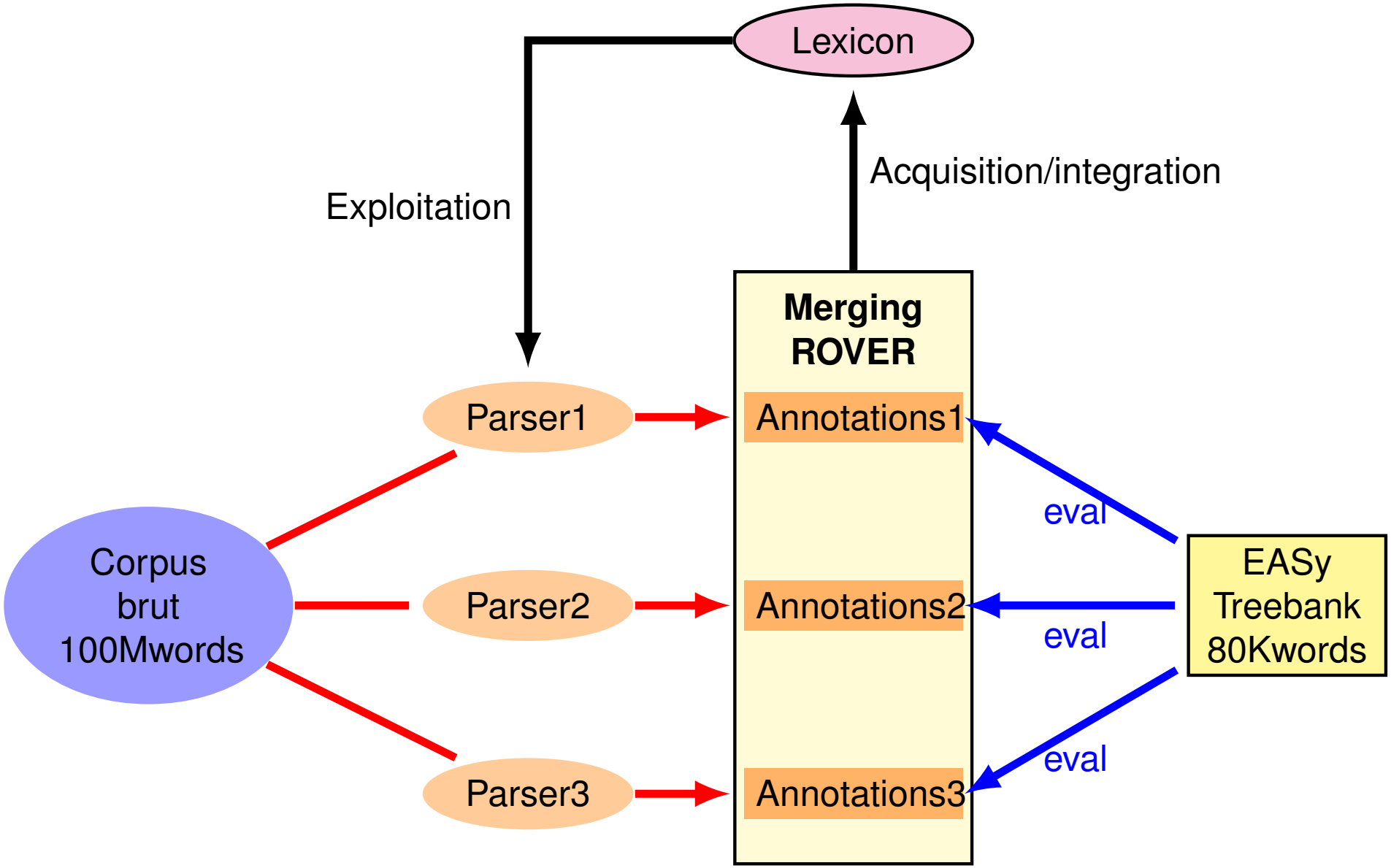
Entering a virtuous loop between tools and resources



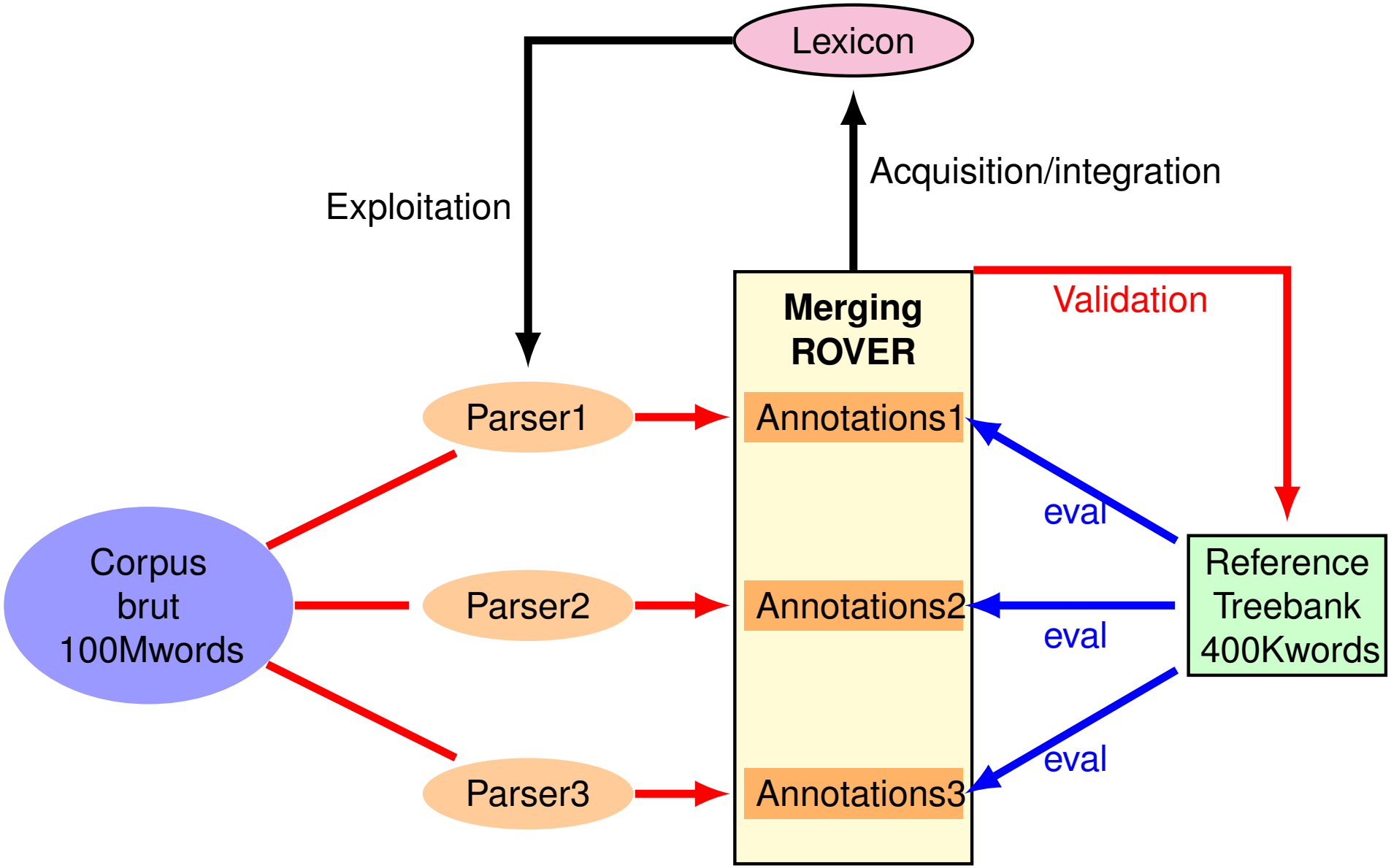
Entering a virtuous loop between tools and resources



Entering a virtuous loop between tools and resources



Entering a virtuous loop between tools and resources



Consortium

ALPAGE
INRIA Paris7

TALARIS/LORIA

LIC2M/CEA-LIST

LIR/LIMSI

Consortium

ELDA

TAGMATICA

Tagmatica

ALPAGE
INRIA Paris7

LPL

TALARIS/LORIA

LIC2M/CEA-LIST

SYNAPSE

LIR/LIMSI

XRCE

LIRMM

10 parsers cooperating

An unique opportunity, source of diversity (formalisms, technologies, . . .)

Parsers	From	Nature
FRMG	INRIA	TIG/TAG+DyALog
SxLFG	INRIA	LFG+Syntax
LLP2	LORIA	TAG
LIMA	CEA-LIST	Rule system
TagParser	TAGMATICA	Induction + rules
GP1 & GP2	LPL	Property grammars
Cordial	SYNAPSE	Rule-based
SYGMART	LIRMM	
XIP	XRCE	Rule-based cascade

Exploiting large corpora

Treebanks are very valuable for NLP but rare and costly to develop.

On the other hand, easy to access large amount of electronic French documents :

Corpus	Size	Type
EASy Corpus	1Mwords	multi-styles
Wikipedia Fr	$\sim$ 86Mwords	collaborative encyclopedia
Wikisources	$\sim$ 80Mwords	free literacy
Monde Diplomatique	18Mwords	journalistic
FRANTEXT	20Mwords	free literacy
Europarl	28Mwords	European Parliament debates
JRC-Acquis	39Mwords	European Law
Corpus Ester	1Mwords	Speech transcription
Total (current)	$>$ **270 Mmots**	

EASy annotations

Based on 6 kinds of chunks and 14 kinds of dependencies

EASy annotations

Based on 6 kinds of chunks and 14 kinds of dependencies

Type	Explanation
GN	Nominal Chunk
NV	Verbal Kernel
GA	Adjectival Chunk
GR	Adverbial Chunk
GP	Prepositional Chunk
PV	Prep. Verbal Kernel

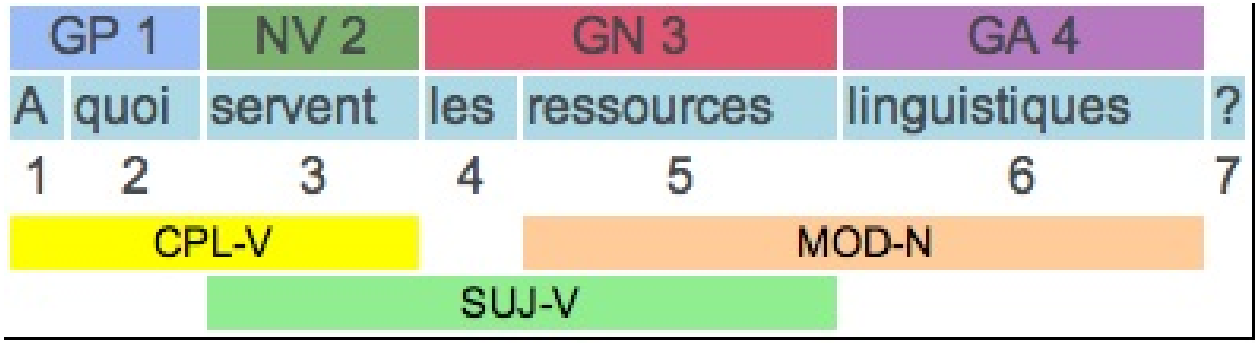
EASy annotations

Based on 6 kinds of chunks and 14 kinds of dependencies

Type	Explanation
GN	Nominal Chunk
NV	Verbal Kernel
GA	Adjectival Chunk
GR	Adverbial Chunk
GP	Prepositional Chunk
PV	Prep. Verbal Kernel

Type	Anchors	Explanation
SUJ-V	suject,verb	Subject-verb dep.
AUX-V	auxiliary, verb	Aux-verb dep.
COD-V	object, verb	direct objects
CPL-V	complement, verb	other verb complements
MOD-V	modifier, verb	verb modifiers
COMP	complementizer, verb	subordinate sentences
ATB-SO	attribute, verb	verb attribute
MOD-N	modifier, noun	noun modifier
MOD-A	mod., adjective	adjective modifier
MOD-R	mod., adverb	adverb modifier
MOD-P	mod., prep.	prep. modifier
COORD	coord., left, right	coordination
APPOS	first, second	apposition
JUXT	first, second	juxtaposition

EASy annotations (cont'd)



Evaluating the parsers

Expertise of **EASy** :

- **[EASy]** End of 2004 (1week)
  Best results (f-measure) : Chunks $>$ 80% Dependencies $>$ 50%

- **[Passage1]** Fall 2007 (2months, closed on Dec. 21st 2007)
  Best results (f-measure) : Chunks $>$ 90% Dependencies $>$ 60%
  - Objective : calibrate the ROVER

- **[Passage2]** End of 2009
  - To be done on new data (from the reference treebank)
  - Objective : To assess the evolutions of the parsers

Data sets

EASy corpus (1Mw)
40K tokenized sentences

journalistic, literacy,
oral, mail, medical, ...

Data sets

EASy corpus (1Mw)
40K tokenized sentences

journalistic, literacy,
oral, mail, medical, ...

easydev (76Kw)
4K annotated sentences
known to participants

Data sets

EASy corpus (1Mw)
40K tokenized sentences

journalistic, literacy,
oral, mail, medical, ...

easydev (76Kw)
4K annotated sentences
known to participants

easytest
400 new annotated sentences

Data sets

EASy corpus (1Mw)
40K tokenized sentences

journalistic, literacy,
oral, mail, medical, ...

easydev (76Kw)
4K annotated sentences
known to participants

easytest
400 new annotated sentences

passagedev (900Kw)
un-tokenized text

wikipedia
wikinews
wikibooks
europarl
jrc-acquis
ester
lemonde

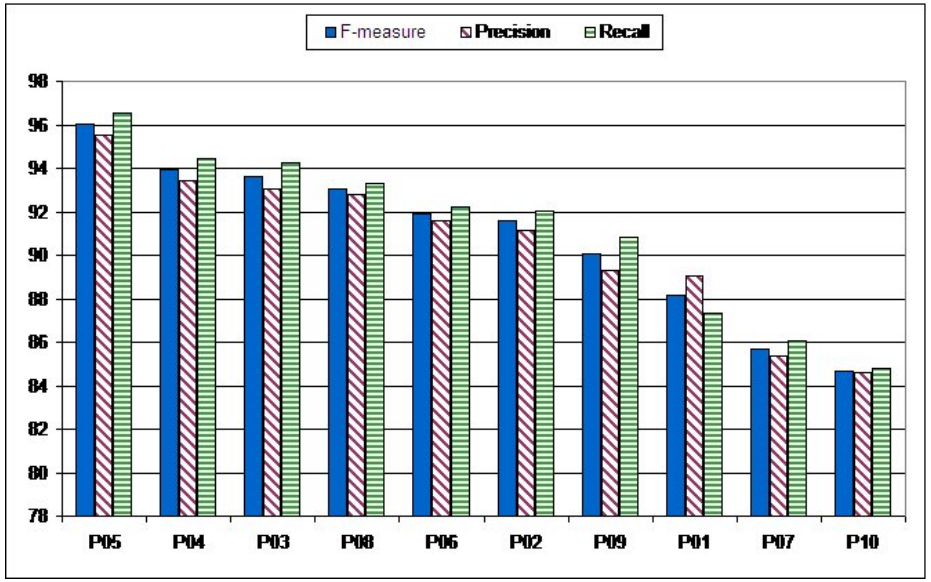
WEB-based evaluation server

- Use of a WEB-based evaluation server
  - Centralized information/data
  - Allow multiple evaluations
  - Instant feedback for participants
    precision, recall, f-measure, plots, logs, ...

- Procedure :
  - Server opened for 2 months
  - Participants upload their outputs
  - Each output submitted is evaluated automatically on the **easydev** data set
    $\Longrightarrow$ immediate feedback
  - Results are kept on the server (max. of ten kept)
  - Before the end, each participant selects a primary submission
  - After the closing, access on the server to the results for the primary submission on the **easytest** data set.

- Conclusion : a very positive initiative
  - Participant P5 submitted more than 50 runs, improving f-measure on chunks from 92.5% to 96% in a few weeks
  - $\Longrightarrow$ the server has been re-opened for new submissions

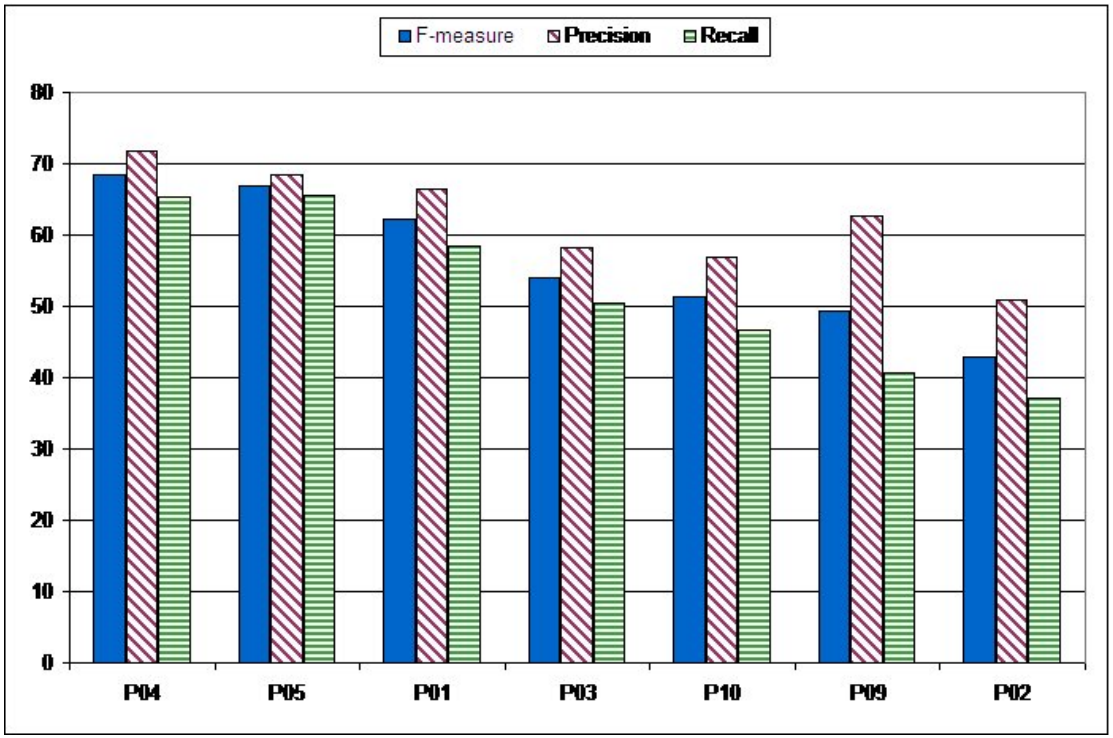
Results on Chunks and Relations (on Test data)



Chunks

- 10 systems
- F-measure :

f	#sys
> 90%	7
> 80%	3

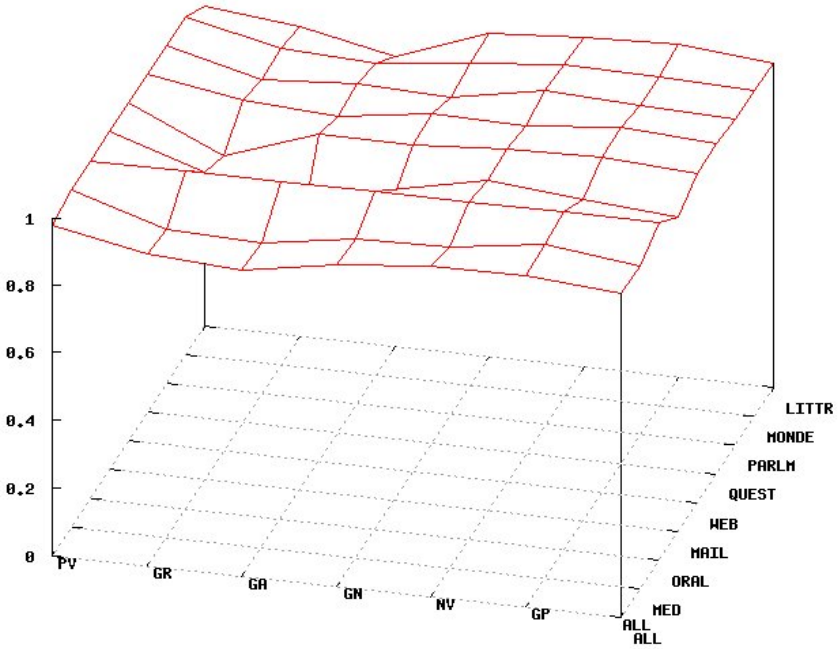


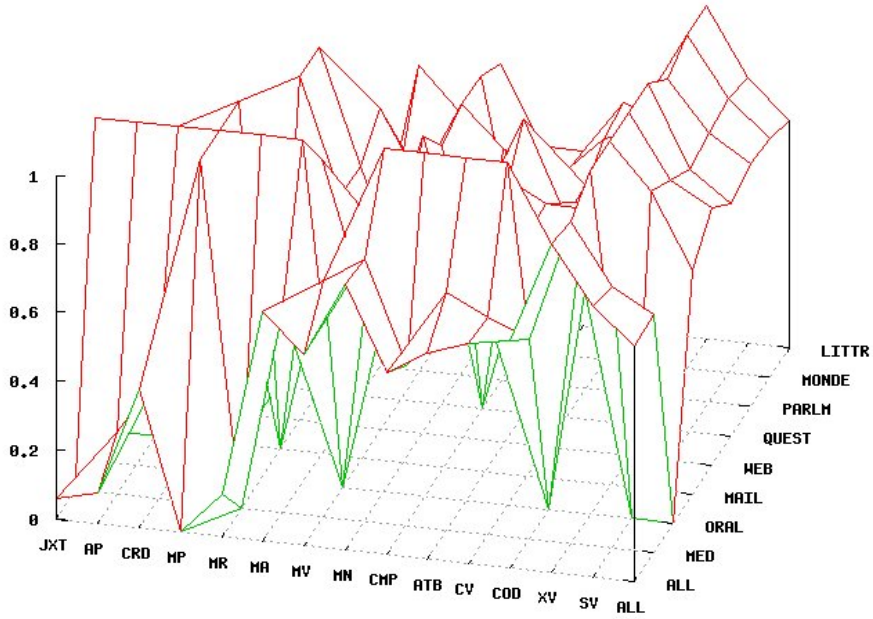
Relations

- 7 systems
- F-measure :

f	#sys
> 60%	3
> 50%	2
> 40%	2

Result landscape (easytest data set)



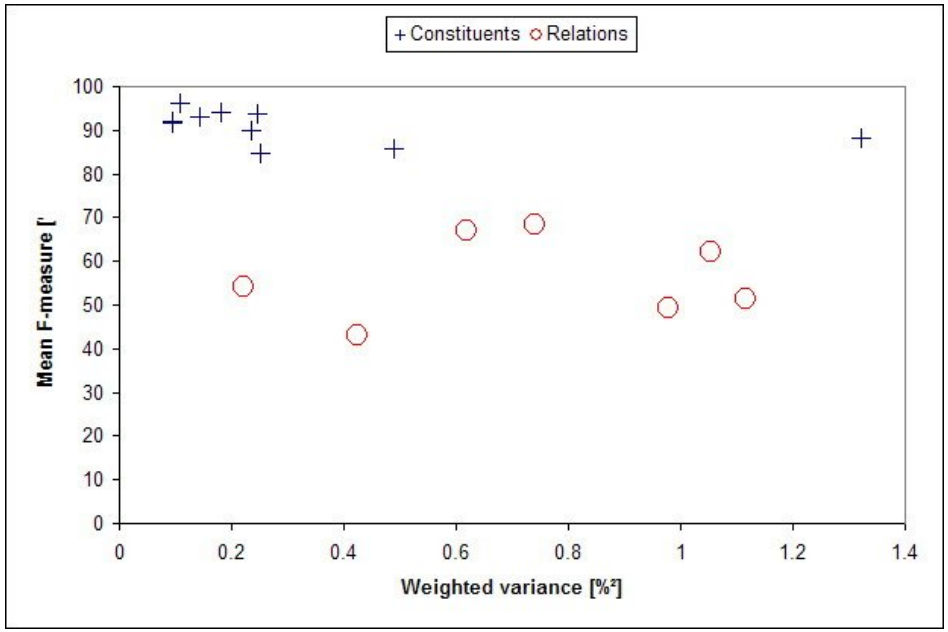


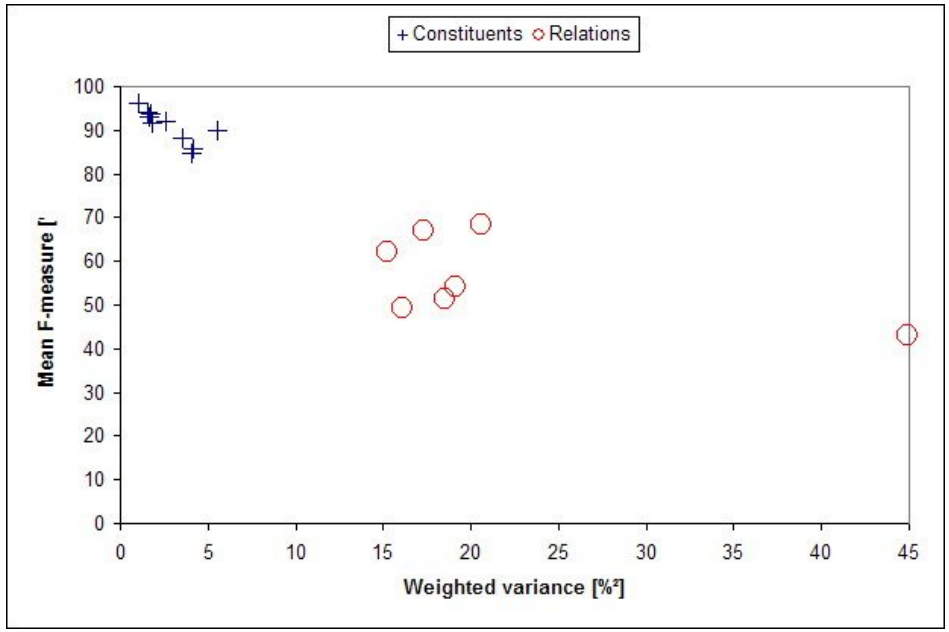
- Performance on Chunks seems very stable wrt corpus and wrt types for this specific system.
- Performances on Relations are less stable, more dependent on relation types

Do we retrieve these properties for the other systems ?

System stability (easytest data set)

Tested using weighted variance





Important to assess the level of stability (confidence) for each system wrt

- corpus type (variances very good, specially for chunks),
- annotation type (larger variances, specially for relations)
- and possibly more specific contexts.

The ROVER : combining the annotations

Using a **ROVER** (*Recognizer output voting error reduction*) to combine the various sets of annotations :

- tried in speech, tagging, translation, ...
- based on majority vote
- pondering parser weights using evaluation results per kind of chunks & relations, corpus style, ...
- feedback on agreement between basic and weighted majority
- control through manual validation
- iterative process

Issues : ensure the coherence of the ROVER annotations

- on chunks : already good results from the parsers and mostly local effects (chunks are small)
- on dependencies : more complex have to enforce dependency properties (single governor, *projectivity*, ...)
- on relationships between chunks and dependencies

ROVER : preliminary algorithm

- apply next steps first on chunks,
  then on relations (constrained by selected chunks)
- confidence for an annotation *a* of type τ in corpus *c*
  produced by participant *i* with rank *r* :

$$\text{conf}^{(i)}(a) = (|\text{systems}| - (r-1)) * \text{prec}^{(i)}_{c,\tau}$$

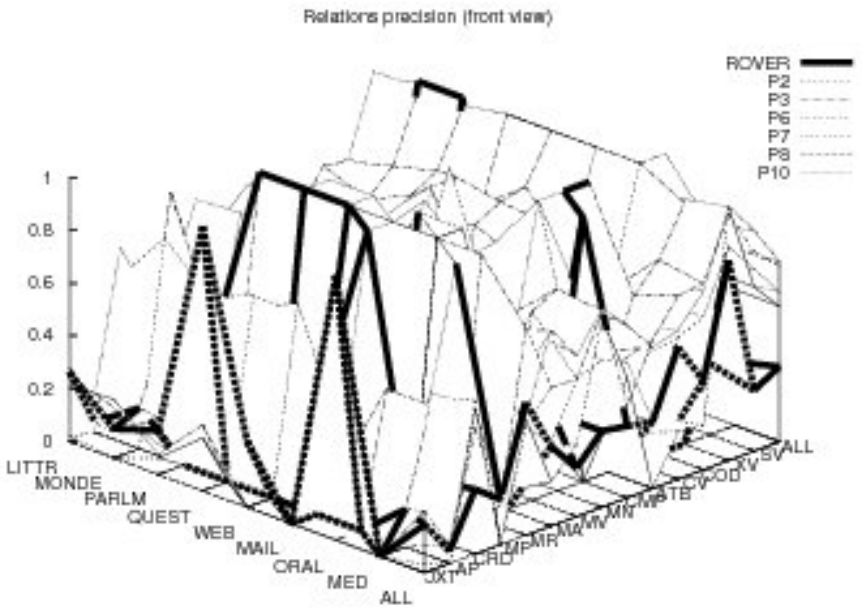
- annotation *a* is selected if

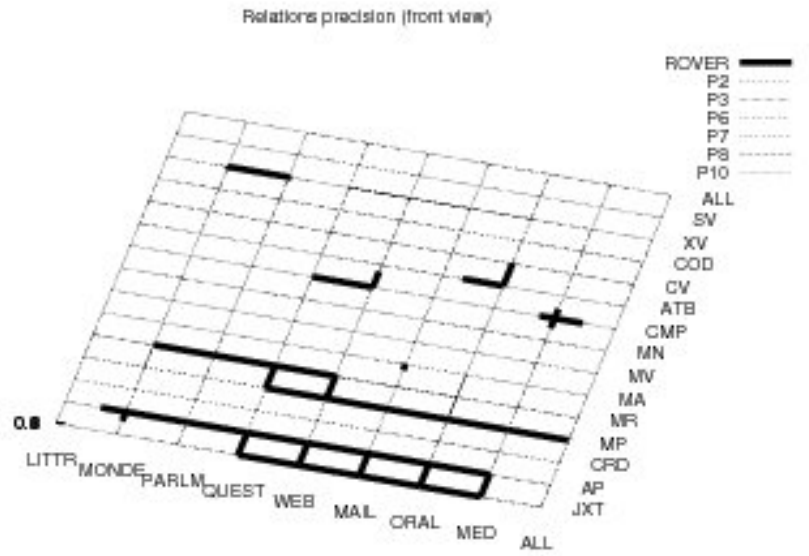
$$p(a) = \frac{\Sigma_i \text{conf}^{(i)}(a)}{|\{i | i \text{ returns } a\}|} \geq \max_i(\text{prec}_{c,\tau})$$

- still very preliminary : many parameters to try
  selection order, confidence weighting, selection threshold, consistence modeling, annotation similarities, ...

- the current algorithm favors precision over recall (maybe to be changed)

ROVER : preliminary results (on 6 participants)





- **bold** : where ROVER has winning precision
- but even when not the best one, not far from the best participant
- and get a confidence level for each annotation

Next steps

- applying machine learning techniques to fine tune the ROVER
  but already encouraging results

- deploying a larger scale infrastructure to work on large corpora
  couping WEB-based evaluation server + ROVER
  + prototype WEB service EASYREF (to view and edit annotations)
  $\Longrightarrow$ iterative collaborative controlled process

- progressively covering the full **Passage** corpus (>100Mwords)

Next steps

- applying machine learning techniques to fine tune the ROVER
  but already encouraging results
- deploying a larger scale infrastructure to work on large corpora
  couping WEB-based evaluation server + ROVER
  + prototype WEB service EASYREF (to view and edit annotations)
  $\Longrightarrow$ iterative collaborative controlled process
- progressively covering the full **Passage** corpus (>100Mwords)

Thank you !