Evaluation of Text Analysis Core Technologies

Two successfull examples : Evaluating POS Taggers and Parsers for French

Patrick Paroubek
Laboratoire pour la Mécanique et les Sciences de l'Ingénieur
Centre National de la Recherche Scientifique

The evaluation paradigm

or the art of tacking pictures

Comparative Evaluation of Technology

- Used sucessfully in the USA by DARPA and NIST (since 1984)
- Similar efforts in Europe on a smaller scale (Sqale, Grace, Senseval, CLEF, Amaryllis, ARC-AUF, Technolangue)
- Select a control task (cf ELSE for definition)
- Gather Participants
- Organize the campaign (protocol/metrics/data)
- Required depending on Technology development stage (with respect to usability/marketability)

Benefits

- Information shared by participants: how to get the best results, as well as access to data
- Information obtained by funding agencies: technology performance, progress/investment, priorities
- Information obtained by industrialists: state of the art, technology choice, market strategy, new products.

Language Resources

- Reference Data manually built (cost + consistency checking + guidelines)
- Definition of Elementary (Linguistic) Data Units
- Quality Criteria
- Language Representativity
- Reutilisability & Multilinguality
- By-products of evaluation (annotated data) become language and evaluation resources

Our two examples: GRACE and EASY

- Comparative evaluation
- Black box evaluation
- Objective evaluation
- Corpus based
- Quantitative measures

GRACE

(the past)

CNRS project

POS tagging?

- Simplest Most Basic Text Analysis Task (Word Classification/Description Nature/Function in Local Context)
- Essential module in many NLP processing (many approaches)
- High performance results
- Common Tagset / Lexicon Problem
- Basic Unit Definition / What's a word?
- Which Metrics?

George Sand a participé a la manifestation.
Tous sont venus l' écouter.

l' is a Pronoun, but with which gender (masculine or feminine)?

Solving POS tagging requires solving the problem of complete Language Understanding (in some cases).

Le programme affiche des résultats.

4 out of the 5 previous words are ambiguous in POS but Contextual Information helps a lot and average POS perplexity is generally located between 1 and 2 (for the main Category).

GRACE, POS Tagging Evaluation for French, 21 participants, 5 countries:

4 phases: training (10 millions words), dry-runs (450.000), tests (836.500), impact study.

17 participants to the dry-run, 13 participants to the final tests

Metrics: precision/decision, measured over 20.000 words, then on 40.000 words with
the EAGLES/ MULTEXT tagset (312 tags)

GRACE

000000 Au DTC:sg
000001 cours SBC:sg
000002 de PREP

Formatting (15 different systems for the tests)

000000 Au Sp+Da-ms-d
000001 cours Ncfs|Ncms
000002 de Da----i|Da-fp-i|Da-mp-i|Sp

Mapping onto GRACE tagset (mapping table provided by participant)

Then align & compare with reference to compute results.

GRACE

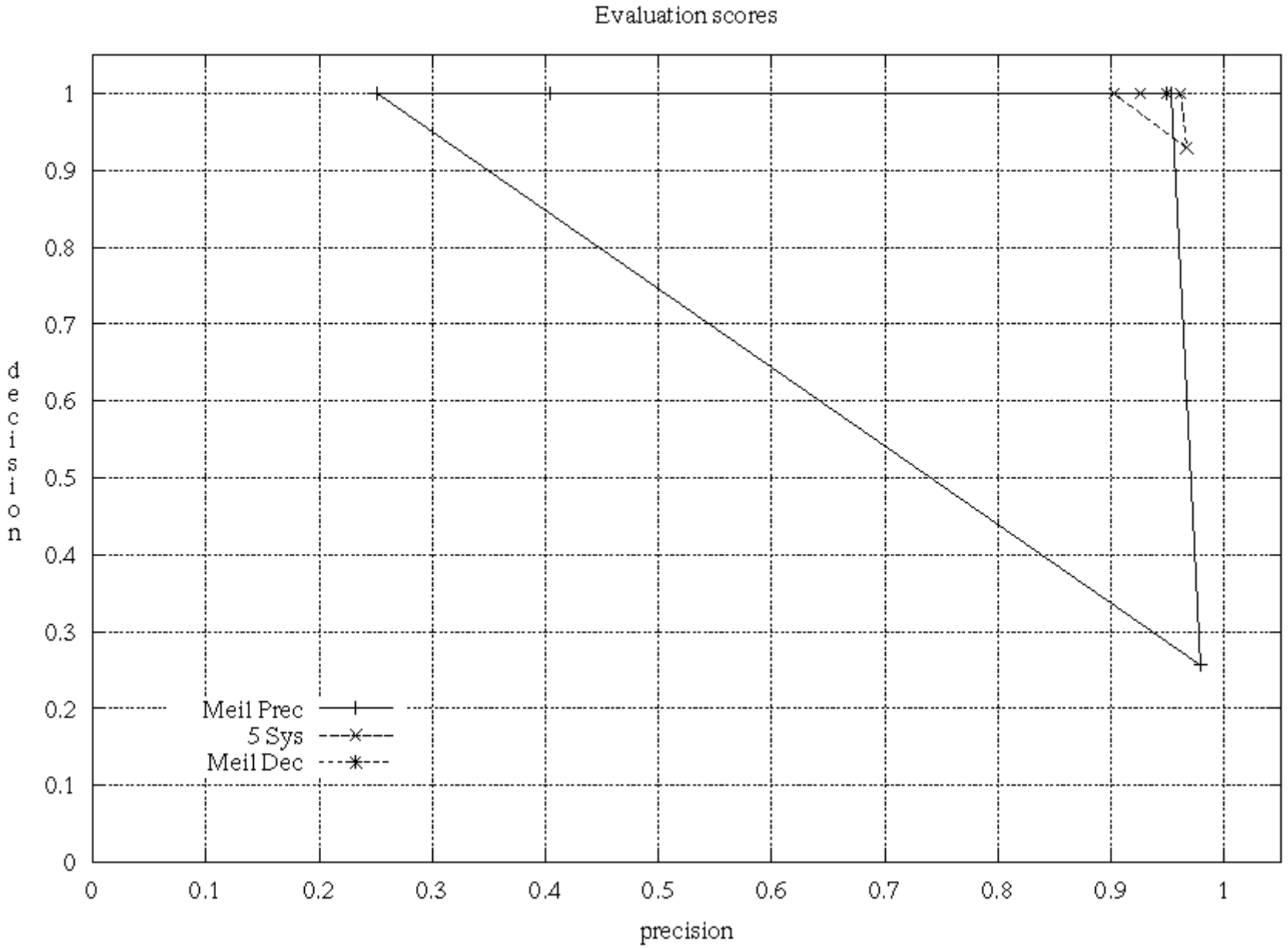
Precision = OK / (OK + ERR)

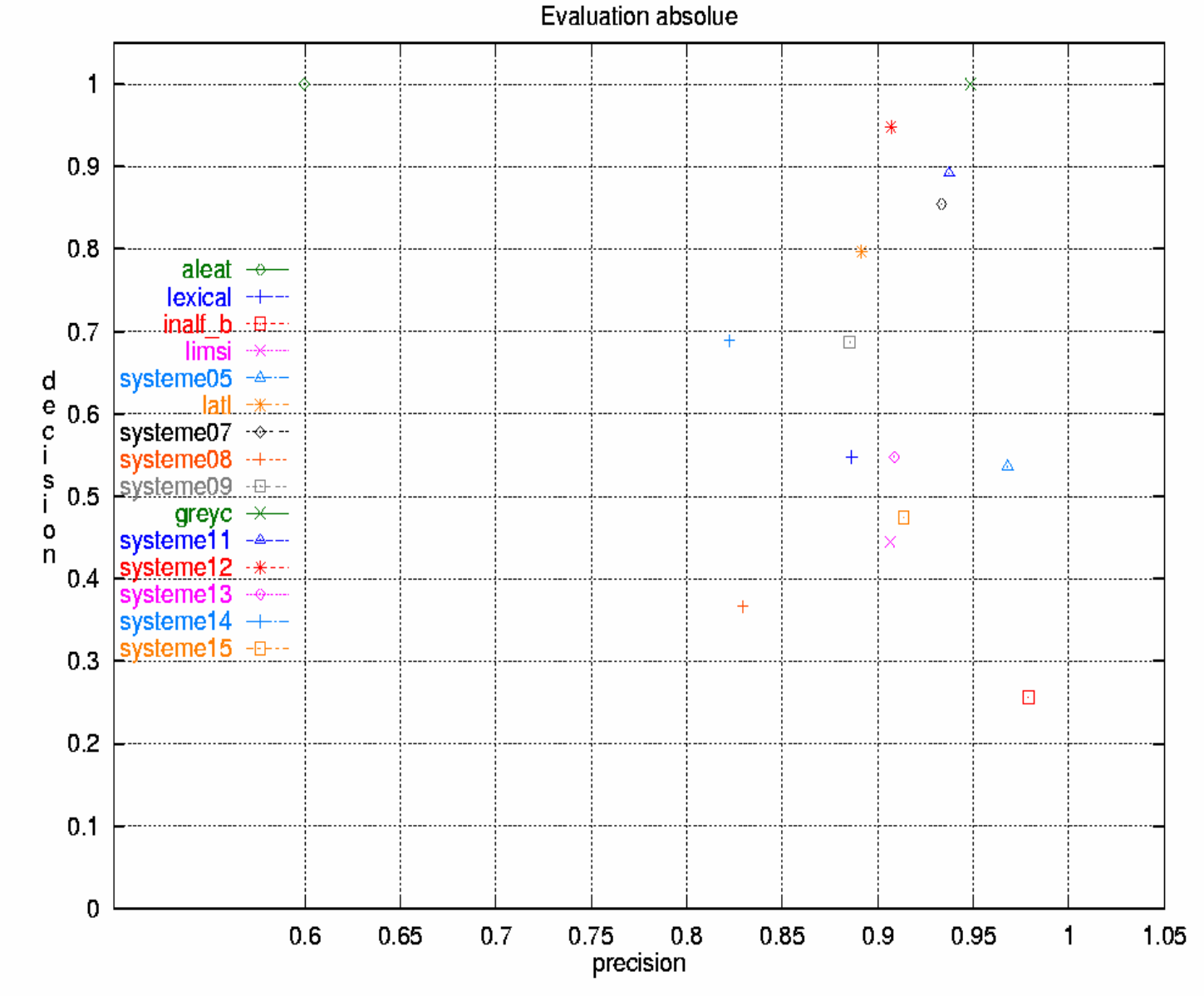
Decision = (OK + ERR) / (OK + ERR + SIL)

OK = nb of forms with 1 correct tag
(full correct disambiguation)

ERR = nb of forms with 1 erroneous tag
(full erroneous disambiguation)

SIL = nb of forms with several tags (partial disambiguation)


Evaluation scores
decision
precision
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1
Meil Prec
5 Sys
Meil Dec


Evaluation absolue
decision
precision
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0.6
0.65
0.7
0.75
0.8
0.85
0.9
0.95
1
1.05
aleat
lexical
inalf_b
limsi
systeme05
latl
systeme07
systeme08
systeme09
greyc
systeme11
systeme12
systeme13
systeme14
systeme15

MULTITAG

Combine to Improve at NIST for Speech Recognition evaluation

ROVER - Recognizer Output Voting Error Reduction (Fiscus 1997)

System combination has better performance than the best system.

Word graph (alignment), majority vote (weighted by maximun occurrence frequency and a confidence score produced by the system).

Error reduction measured by Fiscus: 5,6 % absolute (12,5% relative).

After results combination the data still need to be hand-checked, BUT only on a very small portion of it (less than 10%), and we know which one!

MULTITAG

000000 Au DTC:sg
000001 cours SBC:sg
000002 de PREP

Formatting (15 different systems for the tests)

000000 Au Sp+Da-ms-d
000001 cours Ncfs|Ncms
000002 de Da----i|Da-fp-i|Da-mp-i|Sp

Mapping onto GRACE tagset (mapping table provided by participant)

000000 Au Sp/1.3 6/14[0.428571] 1/4[0.25] 1/14[0.0714286]
000001 cours Ncms|Sp/2.3 6/15[0.4] 1/2[0.5] 3/15[0.2]
000002 de Sp 7/13[0.538462] 1/2[0.5] 4/13[0.307692]

Combination Vote & Confidence Measure

CONCLUSION: GRACE was a success.
Industry and Research met for 5 years on common grounds. As results, a community was created, one participant decided to add a tagger to his product catalog and a new language resource was produced.

GRACE and MULTITAG have proved that the evaluation paradigm can produce high quality validated language resources.

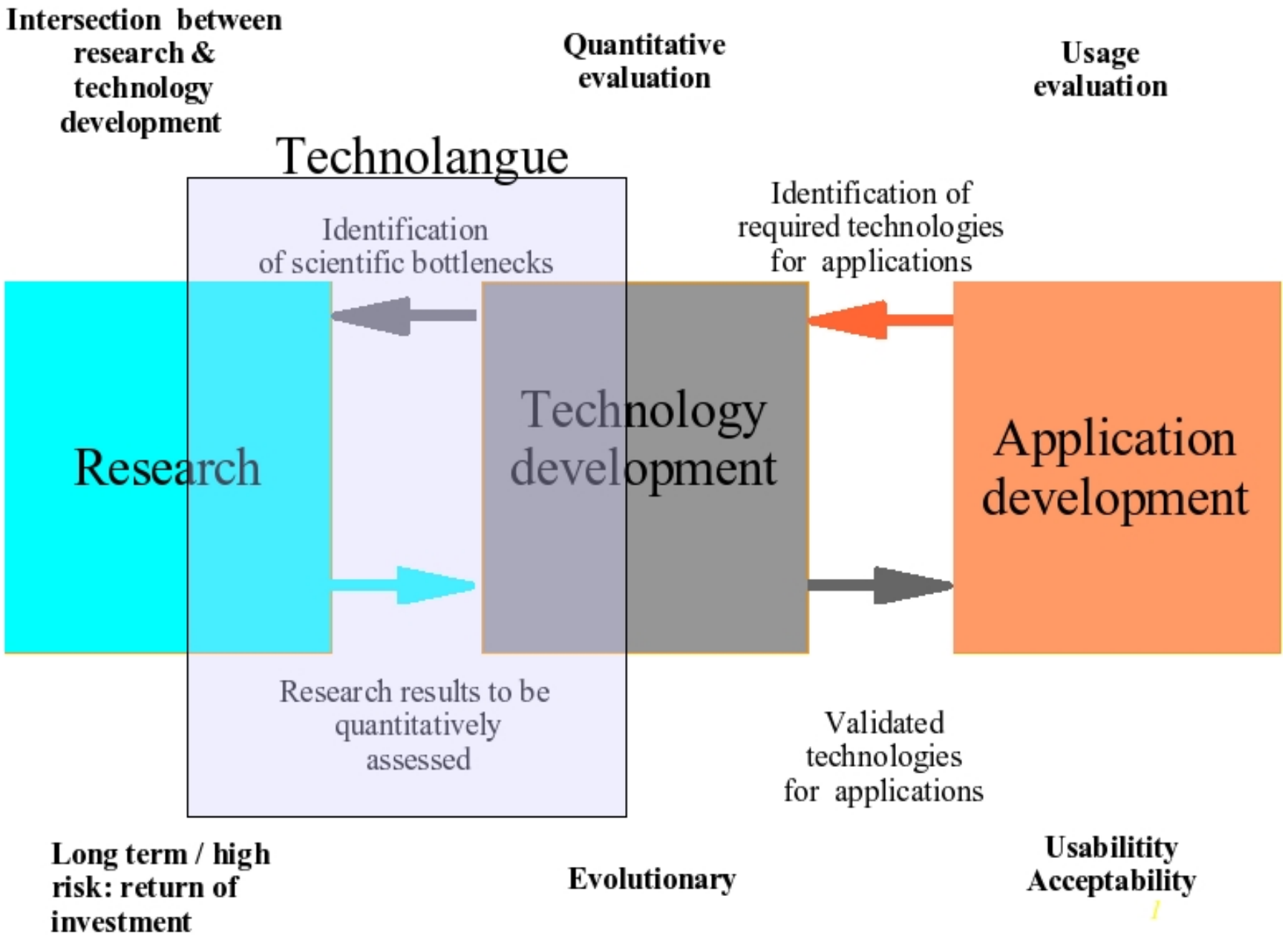
Generalizing this approach to other control tasks could be a mean to increase rapidly and at low cost the amount of annotated and validated language data while deploying the evaluation paradigm.

EASY
(the present)

ELDA-CNRS campaign

in EVALDA

of TECHNOLANGUE


Intersection between research & technology development
Quantitative evaluation
Usage evaluation
Technolangue
Identification of scientific bottlenecks
Identification of required technologies for applications
Research
Technology development
Application development
Research results to be quantitatively assessed
Validated technologies for applications
Long term / high risk: return of investment
Evolutionary
Usabilitity Acceptability

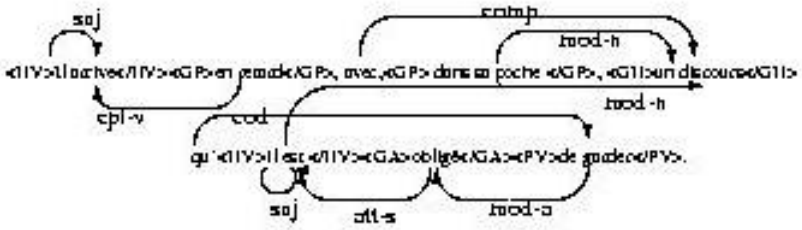
Objective: evaluation of syntactic analysers of French

5 corpous provider, 13 participants, 16 systems

- France Telcom R&D
- GREYC
- INRIA (ATOLL 1,2)
- LATL
- LIC2M
- LIRMM
- LORIA
- XEROX
- LPL (1,2 & 3)
- PERTIMM
- SYNAPSE
- ERSS
- TAGMATICA

Corpus providers :

- ATILF (litterature)
- DELIC (speech transcriptions, emails)
- ELDA (speech ESTER, MLCC, senat, TREC questions
  translated, Amaryllis questions, web)
- LLF (Le Monde)



Il arrive en retard, avec, dans sa poche, un discours qu'il est obligé de garder.

Annotation guide (A. Vilnat) :
http://www.limsi.fr/Recherche/CORVAL/easy/PEAS_reference_annotations_v1.6.html

5 types of constituents

1. GN nominal group
2. GP prepositional group
3. NV verb kernel
4. GA adjectival group
5. GR adverbial group

14 types of relation

1. Subject - Verb
2. Auxiliairy - Verb
3. Direct objet - Verb
4. Complement - Verb
5. Modifier - Verbe
6. Complementer
7. Attribut -Sujet/Objet
8. Modifieur - Nom
9. Modifieur - Adjectif
10. Adverb Modifier
11. Preposition Modifieur
12. Coordination
13. Apposition
14. Juxtaposition

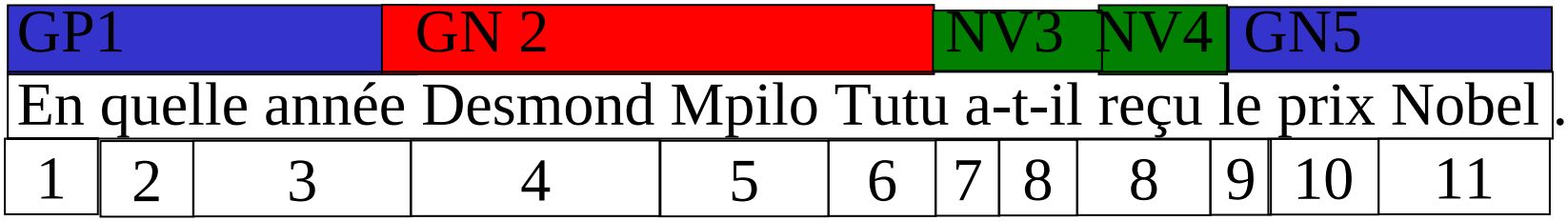
Annotation tool : HTML editor + XML converter (I. Robba)

Manual constituent annotation:

Sentence 1

En quelle année Desmond Mpilo Tutu a-t-il *reçu* le prix *Nobel* …

Sentence 1



Relations

subject	verb
GN2	**F7**
F8	**F7**

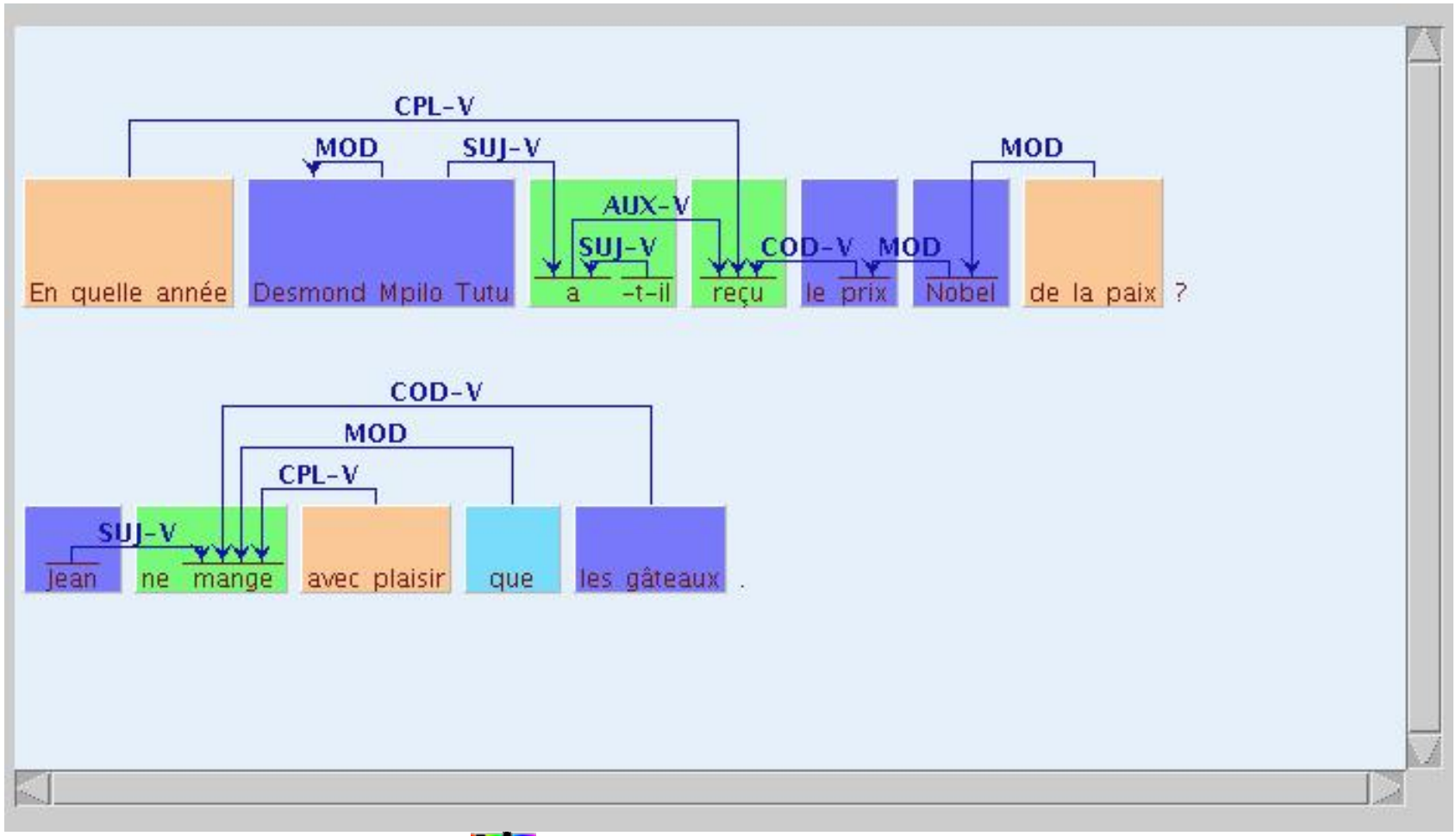
Sentence 12							
NV1			GN2	NV3	GR4	GA5	
Je	pense	que	monsieur	est	très	inquiet	.
1	2	3	4	5	6	7	8

DOC	Verb
NV 3	NV1

Complementer	NV sub. prop
NV 3	NV1

Internal Representation in XML / UTF8 (DTD EASY).

Validation tool : graphic editor (E. Giguet)



Data given to participants :

- Raw
- Segmented into sentences
- Segmented into words and sentences
- Segmented into words and sentences with morphosyntatic annotations (WinBrill + étiquettes GRACE)

Test Corpus annotated by the participants :

769 154 forms 40 260 sentences

Measure Corpus :

83 925 formes 4 269 énoncés

Genre	Test Corpus		Measure Corpus	
	Formes	**Enoncés**	**Formes**	**Enoncés**
Web	16 786	836	**2 104**	**77**
Newspaper	86 273	2 950	**10 081**	**380**
Parliament	81 310	2 818	**8 875**	**298**
Litterature	229 894	8 062	**24 236**	**881**
email	149 328	7 976	**9 243**	**852**
medical	48 858	2 270	**11 799**	**554**
speech	8 106	522	**8 106**	**522**
speech	97 053	11 298	**5 365**	**502**
Questions	51 546	3 528	**4 116**	**203**

Sentences are identified using the typography with regular expressions.

Word forms are defined by regular expression and compounds are given in a list (only function words)

Segmentation of speech DELIC data has been done by hand.

All other data have been segmented using EASY tools.

```
<?xml version="1.0" encoding="UTF-8"?>
<DOCUMENT fichier="\Oral Elda\oral_elda_1EASY.UTF8.xml" xmlns:xlink="http://www.w3.org/1999/xlink">
<E id="E1">
<constituants>
<Groupe type="GN" id="E1G1">
  <F id="E1F1">14</F>
  <F id="E1F2">heures</F>
</Groupe>
<Groupe type="GP" id="E1G2">
  <F id="E1F3">Ã </F>
  <F id="E1F4">Paris</F>
</Groupe>
  <F id="E1F5">,</F>
<Groupe type="GN" id="E1G3">
  <F id="E1F6">midi</F>
</Groupe>
<Groupe type="GP" id="E1G4">
  <F id="E1F7">en</F>
  <F id="E1F8">temps</F>
</Groupe>
<Groupe type="GA" id="E1G5">
  <F id="E1F9">universel</F>
</Groupe>
  <F id="E1F10">,</F>
<Groupe type="GN" id="E1G6">
  <F id="E1F11">l'</F>
  <F id="E1F12">information</F>
</Groupe>
<Groupe type="NV" id="E1G7">
  <F id="E1F13">continue</F>
</Groupe>
```

CONSTITUENTS ANNOTATIONS

<Groupe type="GP" id="E1G8">
<F id="E1F14">sur</F>
<F id="E1F15">RFI</F>
</Groupe>
<F id="E1F16">.</F>
<F id="E1F17">Â§</F>
</constituants>
<relations>
<relation xlink:type="extended" type="MOD-N" id="E1R2">
<modifieur xlink:type="locator" xlink:href="E1G4"/>
<nom xlink:type="locator" xlink:href="E1F6"/>
<a-propager booleen="faux"/>
</relation>
<relation xlink:type="extended" type="SUJ-V" id="E1R3">
<sujet xlink:type="locator" xlink:href="E1G6"/>
<verbe xlink:type="locator" xlink:href="E1G7"/>
</relation>
<relation xlink:type="extended" type="CPL-V" id="E1R4">
<verbe xlink:type="locator" xlink:href="E1G7"/>
<complement xlink:type="locator" xlink:href="E1G8"/>
</relation>
<relation xlink:type="extended" type="MOD-N" id="E1R5">
<modifieur xlink:type="locator" xlink:href="E1G5"/>
<nom xlink:type="locator" xlink:href="E1F8"/>
<a-propager booleen="faux"/>
</relation>
<relation xlink:type="extended" type="MOD-N" id="E1R6">
<modifieur xlink:type="locator" xlink:href="E1F1"/>
<nom xlink:type="locator" xlink:href="E1F2"/>
<a-propager booleen="faux"/>
</relation>
</relations>
</E>

ANNOTATING RELATIONS

Precision-Recall measures :

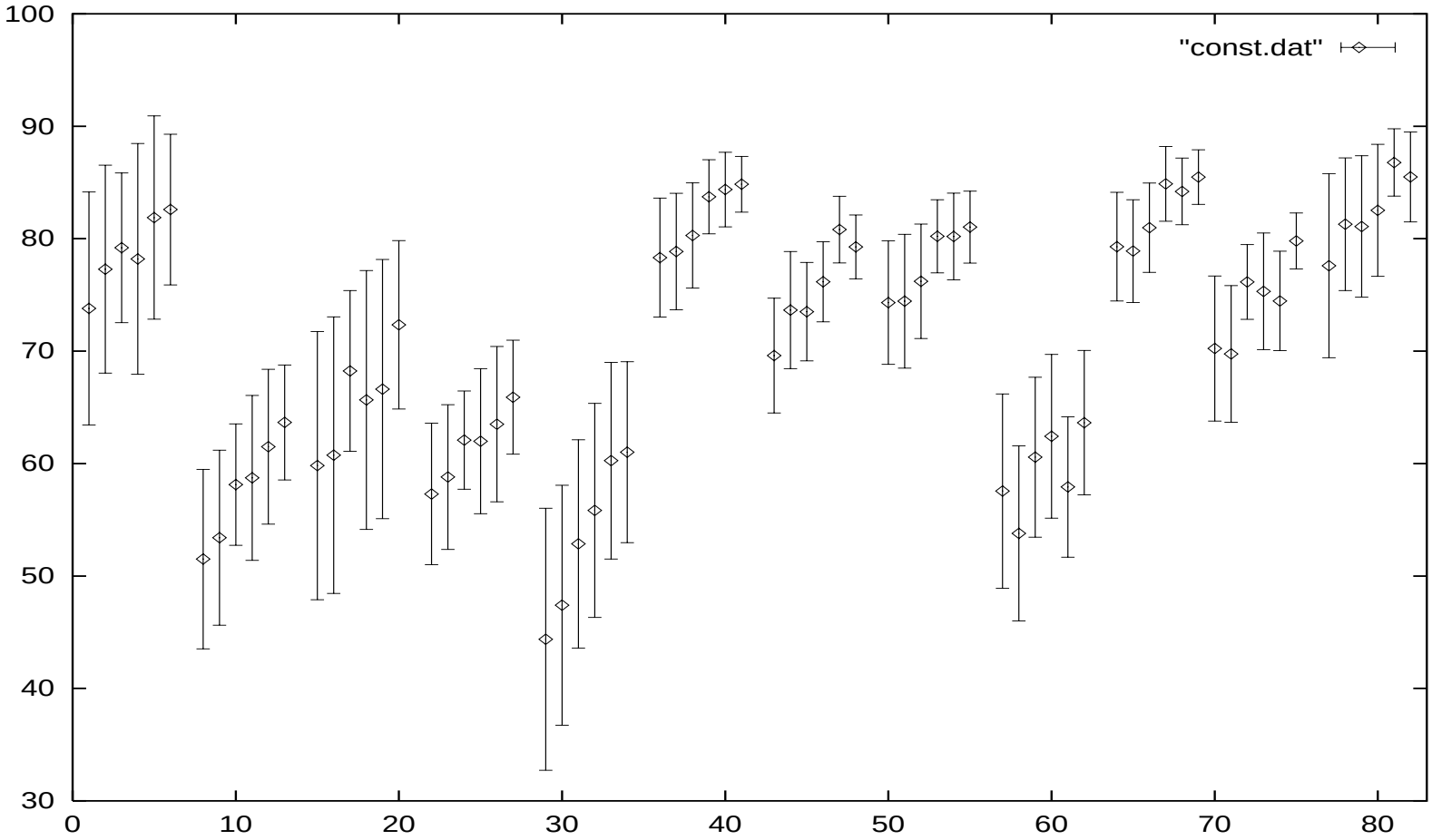
- by participant,
- by type of constituent,
- by type of corpus.

Two modes for measurement:

1) strict measure (equality of word form addresses) and
2) relaxed measure (variation allowed on on beginning and end of group addresses +/-1).

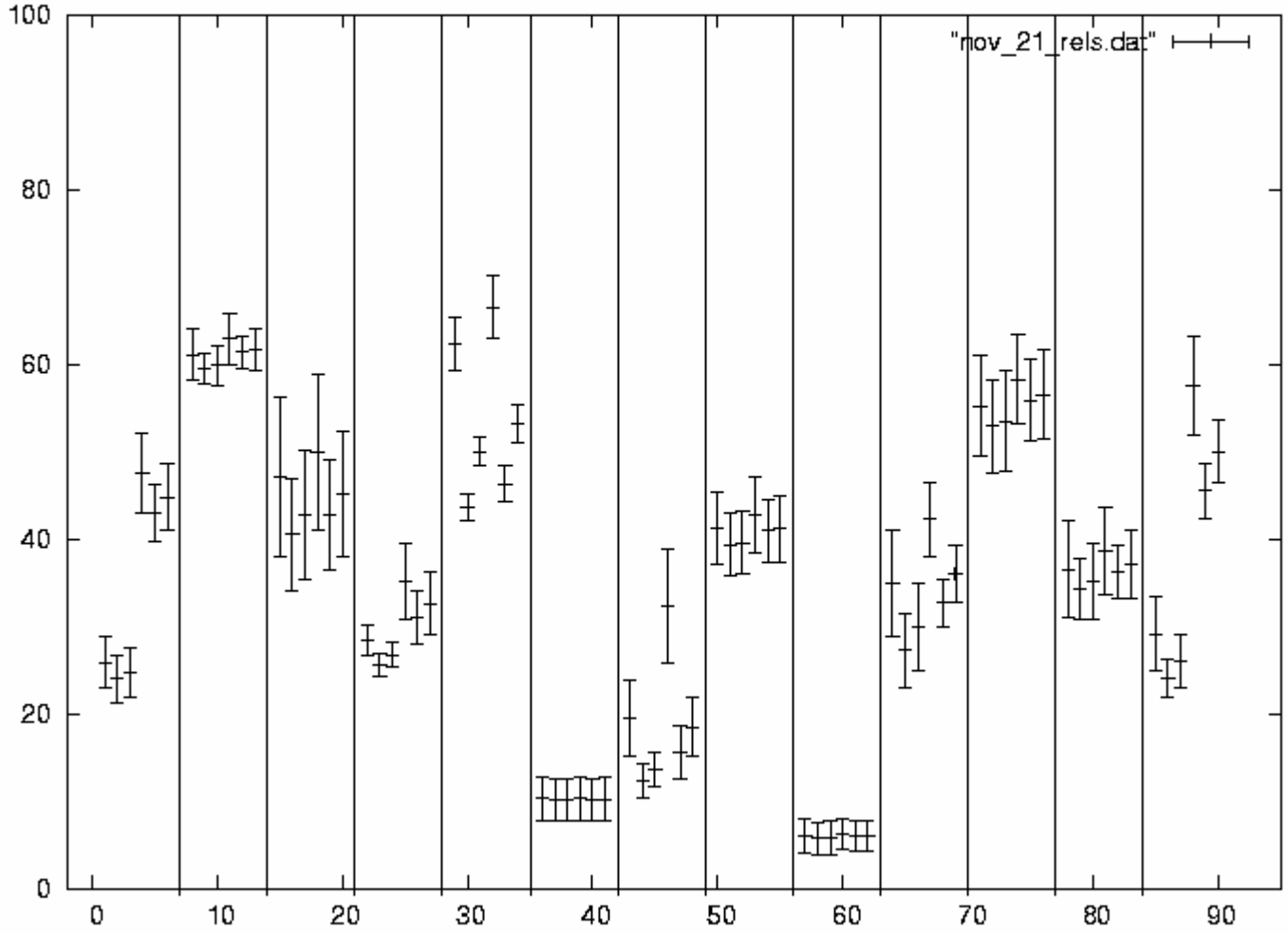
Some surgeneration of relation in reference data for :

1) intra group noun modifier relation (noun-adjective)
2) chained coordinations



Evaluation of constituents for 12 systems
(prec., rec., f-mes., and the same in relaxed mode)

Evaluation in relations for parliament, senat and litteraire_1 for 13 systems.



CONCLUSION:

Although the task is much more difficult than for GRACE, EASY is also on a path to success.

Industry and Research have been meeting now for 3 years on common grounds.

As results, a community was created, that agreed on a common format, annotations and evaluation metrics...

What next ?

international campaign ?

European campaign ?