

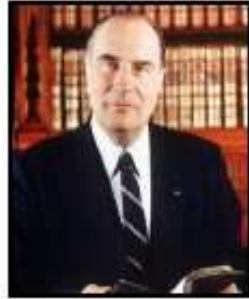


DEFT'05

DÉfi francophone Fouille de Textes

Jérôme Azé et Mathieu Roche
Responsables du Comité d'Organisation de DEFT'05

Violaine Prince et Yves Kodratoff
Co-Présidents du Comité de Programme de DEFT'05



DEFT'05
DÉfi Fouille de Textes

Atelier de TALN'05
10 juin 2005, 14h-17h30,
Dourdan (91)



- Insertion de passages de discours de F. Mitterrand dans les discours de J. Chirac.
- **But du défi** : identifier les passages de F. Mitterrand introduits.
- Défi proche de la tâche Novelty de TREC.
- Trois tâches (*3 exécutions maximum par tâche*)
 - **Tâche 1** : corpus sans dates ni noms de personnes.
 - **Tâche 2** : corpus sans dates.
 - **Tâche 3** : corpus avec dates et noms de personnes.

Préparation des données (1/4)

- Acquisition des corpus
 - Discours de J. Chirac : <http://elysee.fr>
 - Discours de F. Mitterrand : <http://discours-publics.ladocumentationfrancaise.fr>
- Normalisation des corpus
 - Suppression des balises HTML, des en-têtes des discours, conversion des entités SGML en caractères ISO8859-1.
 - Placer une phrase par ligne (traiter les points relatifs aux abréviations tel que "M." pour "Monsieur")
 - ...

Préparation des données (2/4)

- Expertise des corpus
 - **Catégorisation des discours** : national (36.6%), international (47.2%) et mixte (16.2%).
- Introduction des phrases de F. Mitterrand dans le corpus de J. Chirac
 - **Croisement des thématiques** nationales et internationales.
 - **Sélection des extraits** de discours de F. Mitterrand les plus "proches" de J. Chirac.
 - **Introduction d'au plus un passage** de F. Mitterrand dans chaque discours de J. Chirac.

Préparation des données (3/4)

- Introduction des phrases de F. Mitterrand les "plus proches" dans le corpus de J. Chirac :

$$score(d_C^{cat}, p_M^{\overline{cat}}) = score_{car}(d_C^{cat}, p_M^{\overline{cat}}) + score_{mot}(d_C^{cat}, p_M^{\overline{cat}})$$

cat : international ou national

d_C^{cat} : discours de J.Chirac appartenant à *cat*

p_M^{cat} : partie de discours de F.Mitterrand appartenant à *cat*

- Déterminer les 20 meilleurs $p_M^{\overline{cat}}$ tels que $score(d_C^{cat}, p_M^{\overline{cat}})$ soit maximum.
- Insertion aléatoire du "meilleur" passage de F. Mitterrand non utilisé.

Préparation des données (4/4)

- Identification des dates et noms de personnes pour constituer les corpus des tâches 1 et 2 :
 - **Années** comprises entre 1900 et 2099
 - **Noms de personnes** : couples de mots commençant par une majuscule et avec éventuellement une particule intercalée, particules suivies d'un mot en majuscule et noms en majuscules isolés. Ces noms ont été normalisés.

**Merci aux membres du Comité d'Organisation
pour la préparation des données !!**

Erick Alphonse (INRA - MIG)

Ahmed Amrani (ESIEA & LRI - IA)

Jérôme Azé (LRI - IA)

Thomas Heitz (LRI - IA)

Amar-Djalil Mezaour (LRI - IASI)

Mathieu Roche (LRI - IA)

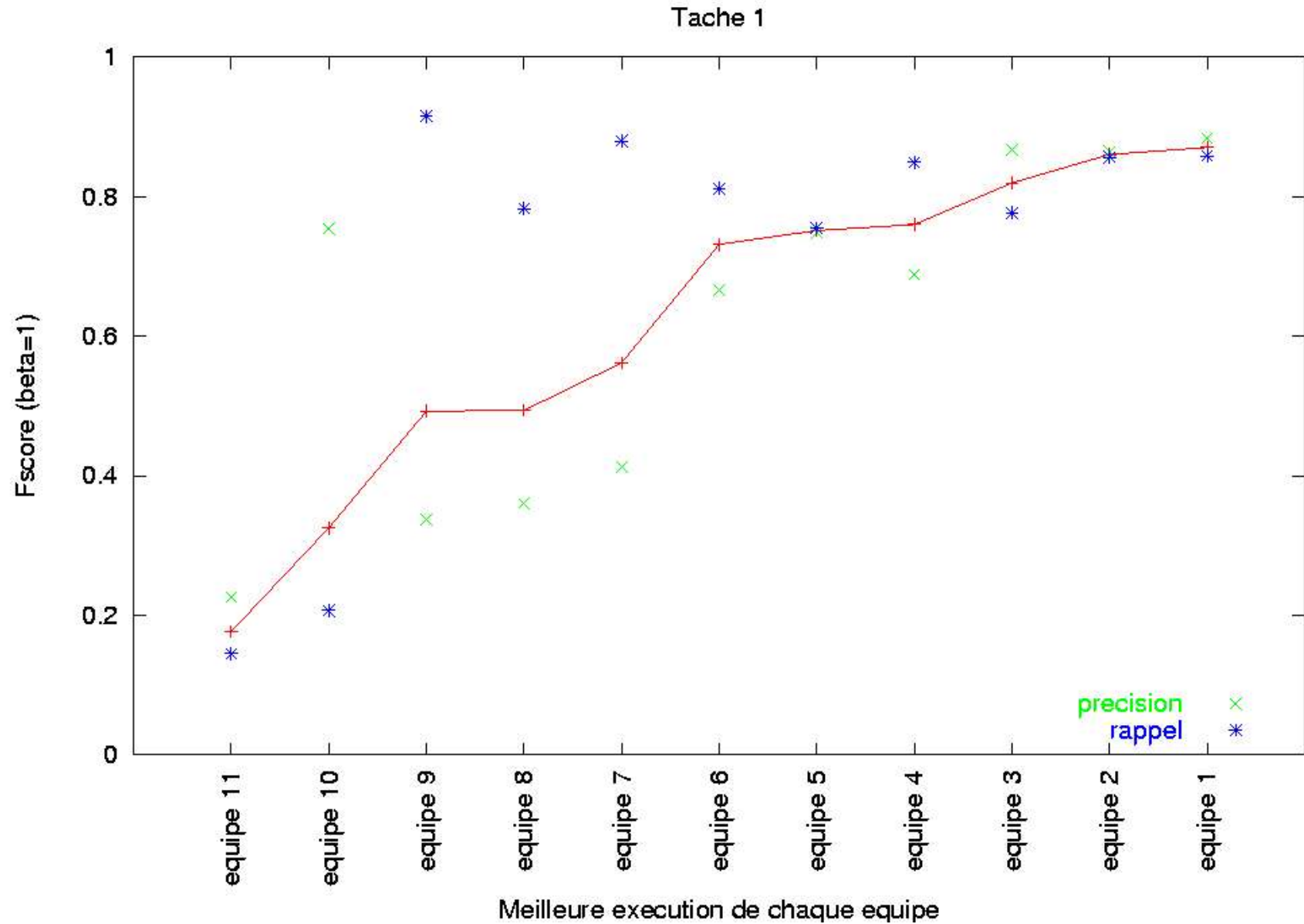
Critères d'évaluation

- **Précision** : pourcentage de phrases correctement associées à F. Mitterrand dans le fichier résultat parmi les phrases soumises.
- **Rappel** : pourcentage de phrases correctement associées à F. Mitterrand dans le fichier résultat parmi les phrases de F. Mitterrand réellement introduites dans le corpus.
- F score ($\beta = 1$) :
$$F_{score}(\beta) = \frac{(\beta^2 + 1) \times \text{Précision} \times \text{Rappel}}{\beta^2 \times \text{Précision} + \text{Rappel}}$$

Résultats des meilleures exécutions

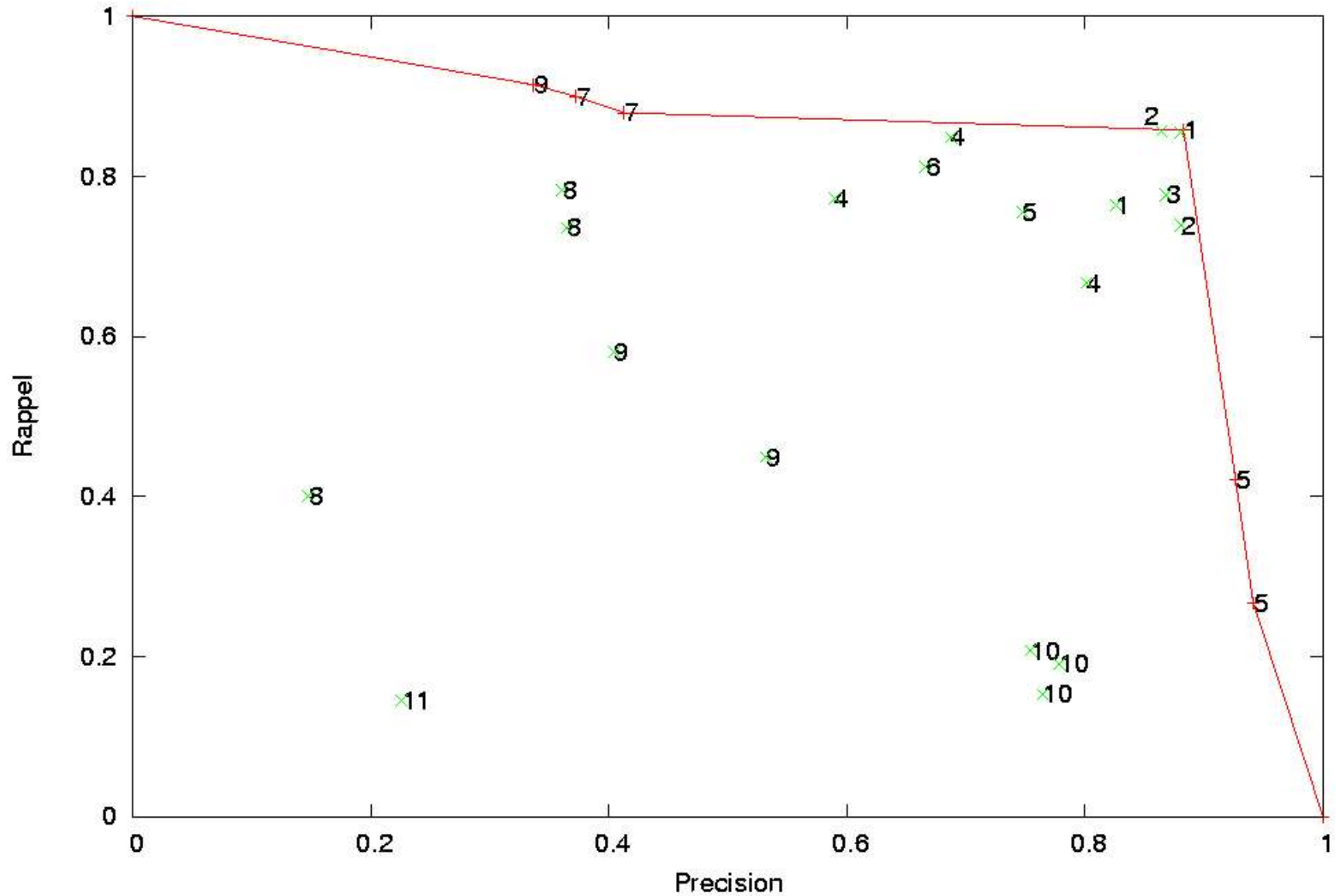
| | tâche 1 | tâche 2 | tâche 3 |
|------------------|---------|---------|---------|
| Équipe 1 | 0.87 | 0.88 | 0.88 |
| Équipe 2 | 0.86 | 0.85 | 0.87 |
| Équipe 3 | 0.82 | 0.82 | 0.82 |
| Équipe 4 | 0.76 | 0.74 | 0.75 |
| Équipe 5 | 0.75 | 0.75 | 0.76 |
| Équipe 6 | 0.73 | 0.79 | 0.79 |
| Équipe 7 | 0.56 | 0.56 | 0.57 |
| Équipe 8 | 0.49 | 0.52 | 0.51 |
| Équipe 9 | 0.49 | 0.56 | 0.56 |
| Équipe 10 | 0.32 | 0.31 | 0.31 |
| Équipe 11 | 0.18 | 0.18 | 0.42 |

F score (tâche 1)



Front de Pareto (tâche 1)

Tache 1





Le prix DEFT'05 est attribué à l'équipe suivante :

*Frédéric Béchet
Marc El-Bèze
Juan-Manuel Torres-Moreno*



CLASSEMENT GLOBAL (*meilleure soumission de chaque équipe*) :

Classement tâches 1 :

1^{er} : Équipe de *F. Béchet, M. El-Bèze, J.-M. Torres-Moreno* - LIA

2^{ème} : Équipe de *O. Cappé, L. Rigouste, F. Yvon* - ENST

3^{ème} : Équipe de *C. Durkal, S. Freydiger, L. Pierron* - LORIA

Classement tâches 2 :

1^{er} : Équipe de *F. Béchet, M. El-Bèze, J.-M. Torres-Moreno* - LIA

2^{ème} : Équipe de *O. Cappé, L. Rigouste, F. Yvon* - ENST

3^{ème} : Équipe de *C. Durkal, S. Freydiger, L. Pierron* - LORIA

Classement tâches 3 :

1^{er} : Équipe de *F. Béchet, M. El-Bèze, J.-M. Torres-Moreno* - LIA

JEUNES CHERCHEURS :

Classement tâches 1 :

1^{er} : Équipe de *A. Labadié*

Y. Romero, L. Sitbon - LIA

2^{ème} : Équipe de - *L. Maisonnasse,*

C. Tambellini - CLIPS

Classement tâches 2 :

1^{er} : Équipe de - *L. Maisonnasse,*

C. Tambellini - CLIPS

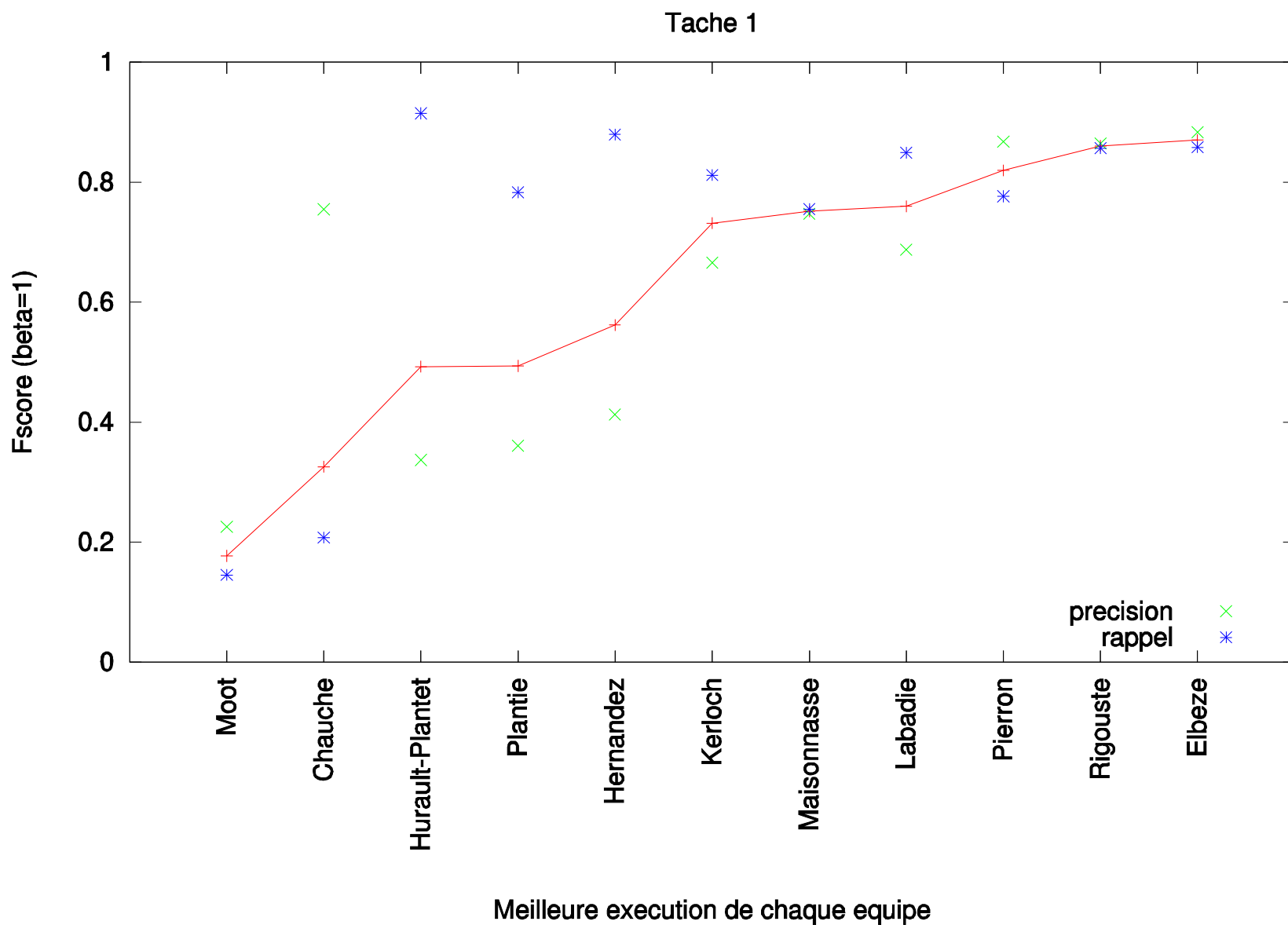
2^{ème} : Équipe de *A. Labadié*

Y. Romero, L. Sitbon - LIA

Classement tâches 3 :

1^{er} : Équipe de - *L. Maisonnasse,*

F score (tâche 1)



Résultats des meilleures exécutions

| | tâche 1 | tâche 2 | tâche 3 |
|------------------------------|---------|---------|---------|
| Elbeze_LIA | 0.87 | 0.88 | 0.88 |
| Rigouste_ENST | 0.86 | 0.85 | 0.87 |
| Pierron_LORIA | 0.82 | 0.82 | 0.82 |
| Labadie_LIA | 0.76 | 0.74 | 0.75 |
| Maisonnasse_CLIPS | 0.75 | 0.75 | 0.76 |
| Kerloch_LIP6 | 0.73 | 0.79 | 0.79 |
| Hernandez_LIMSI | 0.56 | 0.56 | 0.57 |
| Plantie_LGI2P | 0.49 | 0.52 | 0.51 |
| Hurault-Plantet_LIMSI | 0.49 | 0.56 | 0.56 |
| Chauche_LIRMM | 0.32 | 0.31 | 0.31 |
| Moot_LABRI | 0.18 | 0.18 | 0.42 |