Préparation et organisation de DEFT 2005 et 2006

Thomas Heitz et Jérôme Azé - Séminaire LIMSI LIR

Novembre 2006

Naissance du défi

DEFT : DÉfi Fouille de Textes

- Inspiré de la compétition américaine TREC Novelty.
- Comparer les méthodes de fouille de textes (FDT) francophones.
- Fédérer les équipes de recherches francophones en FDT.
- Mettre des corpus étiquetés à la disposition de la communauté.
- Thématiques : attribution d'auteur (2005), segmentation thématique (2006).

Préparation des corpus

Édition 2005

- Discours politique de J. Chirac et F. Mitterrand : 14 et 17 Mo ;
- Découpage apprentissage / test : 70/30.
- Phases : récupération, conversion, remplacement et suppression, segmentation, fusion, contrôle.
- Contrôle : 5 Mo / personne.
- Durée totale : 3 mois environ.

Préparation des corpus

Édition 2006

- Discours politique de J. Chirac, F. Mitterrand et V-G. d'Estaing : 70 Mo, lois de l'union européenne : 110 Mo et ouvrage scientifique : 1 Mo.
- Découpage apprentissage / test : 60/40.
- Phases : récupération, conversion, remplacement et suppression, segmentation, contrôle.
- Contrôle : 30 Mo / personne.
- Durée totale : 3 mois environ.

Organisation du défi

Sites internet, ateliers, participants

- Soumissions en ligne : 3x3 fichiers (2005), 3 fichiers (2006)
- En 2005, l'AFIA a financé 300€pour le premier prix, inscriptions et actes gérés par TALN. Soumission des articles à la revue RNTI.
- En 2006, 15€inscription atelier en plus de la SDN (85€), le LRI a financé les actes (tirages + CDrom, transports).
- 26 (respectivement 22) personnes présentes pour 10 (7) équipes participantes lors de DEFT 2005 (2006).

Résultats des participants

Mesure d'évaluation

$$F_{score}(\beta = 1) = \frac{(\beta^2 + 1) \times Précision \times Rappel}{\beta^2 \times Précision + Rappel}$$

Les scores et classements

- Front de Pareto ; F_{score} souple (2006)
- Les soumissions sont considérées comme indépendantes. Classement par tâches (2005), classement global par rang sur tous les corpus (2006).
- F_{score} compris entre 0.2 et 0.9 pour 2005 (0 et 0.6 pour 2006), max sur tâches 2 et 3, (max sur les lois et min sur le livre).

Plan

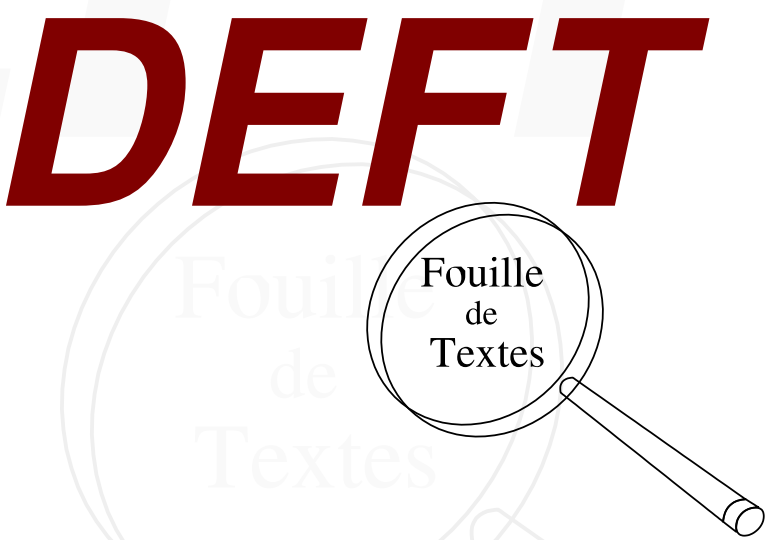
2 Naissance
- Inspiration
- DEFT 2005
- DEFT 2006

3 Préparation

4 Organisation

5 Évaluation et résultats

6 Annexes : résultats



Inspiration : MUC et TREC

Message Understanding Conference Proceedings 1998

- MUC 7 : Information Extraction (Template Element, Template Relation, Scenario Template), Named Entity, Coreference.
- Ancêtre de TREC.

Text REtrieval Conference 2004 - http ://trec.nist.gov

- Tâche Novelty : dans un corpus d'actualités américaines de 10 Mo, trouver les phrases qui correspondent à un paragraphe décrivant un sujet.
- De plus, dire si les phrases trouvées sont nouvelles par rapport à celles précédemment trouvées.

Pourquoi reconnaître les auteurs ?

Motivations de DEFT 2005

- Détection des passages les plus singuliers dans des textes quelconques (rupture de style, changement de contexte).
- Détection de plagiats possibles dans des textes.
- Détection des informations générales dans des corpus techniques.

Tâches proposées

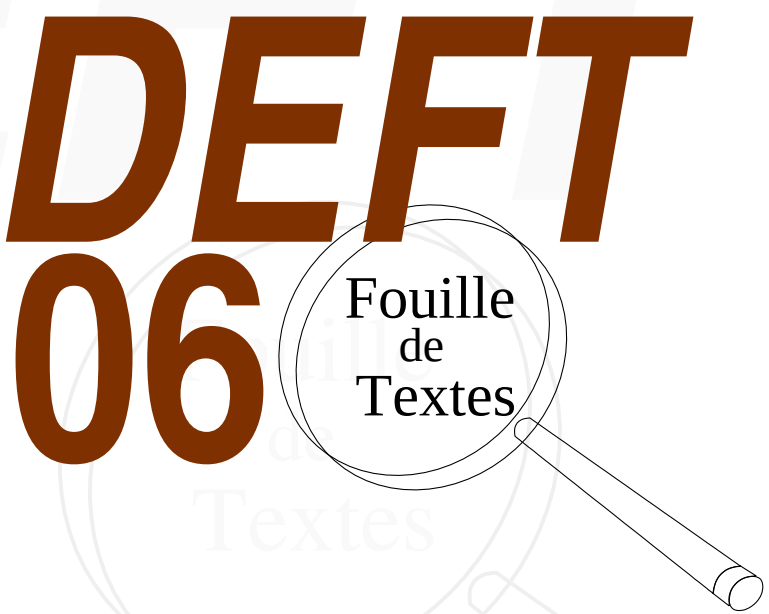
Insertion de passages de F. Mitterrand dans les discours de J. Chirac.

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

Pourquoi segmenter thématiquement ?

Motivations de DEFT 2006

- Délimiter des zones de textes dans des résultats de requêtes.
- Résumer un texte tout en conservant ses principaux thèmes.
- Sélectionner le bon dictionnaire lors d'une correction orthographique ou d'une traduction.



Qu'est ce qu'un segment thématique ?

Segmentation voulue par les auteurs des textes ; la plus simple pour préparer les corpus.

Définition d'un segment

1. **Discours politiques** : division en paragraphes thématiques lors de leur écriture.
2. **Lois européennes** : les segments thématiques sont les lois.
3. **Ouvrage scientifique** : chapitres, sections, sous-sections et sous-sous-sections.

Conclusion sur la naissance

Conclusion

- Adaptation francophone de compétitions américaines.
- Comparer les méthodes de fouille de textes (FDT) francophones.
- Fédérer les équipes de recherches francophones en FDT. 10 (7) équipes participantes lors de DEFT 2005 (2006).
- Création de corpus mis à disposition de la communauté francophone.

Plan

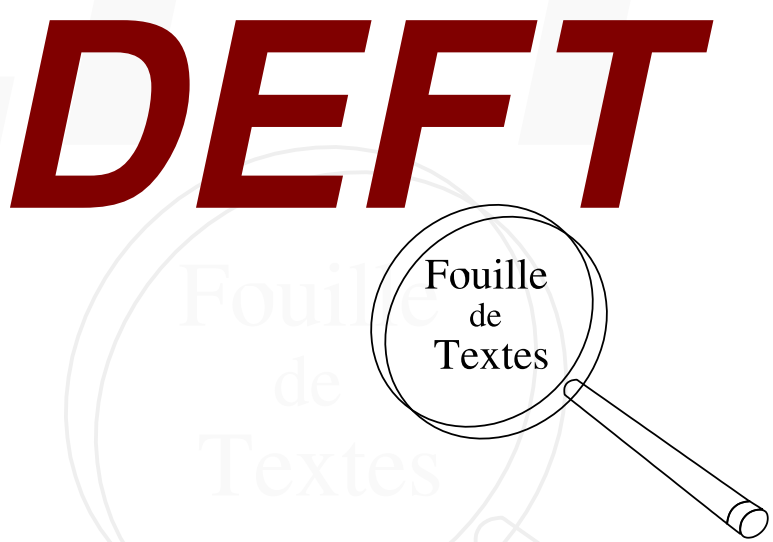
2 Naissance

3 Préparation
- DEFT 2005
- DEFT 2006

4 Organisation

5 Évaluation et résultats

6 Annexes : résultats



Préparation des données (1/4)

Acquisition des corpus

- Discours de J. Chirac : elysee.fr
- F. Mitterrand : 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 eux abréviations tel que "M.") ...

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.
- 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)

Fusion des corpus

- Introduction des passages de F. Mitterrand les "plus proches" dans le corpus de J. Chirac.
- Pour chaque discours de J. Chirac, déterminer les 20 passages les plus proches du discours. Insertion aléatoire du "meilleur" passage de F. Mitterrand jamais utilisé.

Préparation des données (4/4)

Retrait des dates et noms de personnes

- 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.

Pourquoi ces corpus ?

Caractéristiques des corpus

- Lois et discours politiques appartenant au domaine public.
- Tailles, thèmes, styles variés.
- Données réelles et courantes.

Préparation des corpus 1/2

Phases de la préparation des données

1. **Récupération** des données brutes, *ex. discours, téléchargement d'un site internet*
2. **Conversion** au format texte, *ex. ouvrage scientifique, conversion des fichiers depuis LaTeX*
3. **Remplacement et suppression** des parties non textuelles ou inexploitables, *ex. ouvrage scientifique, problème des formules*
4. **Segmentation** thématique, *selon le type de corpus*
5. **Contrôle** qualité, *180 Mo, 6 personnes*

Préparation des corpus 2/2

Spécificités

1. **Discours politiques** : titre, date, orateur supprimés ; discours de V-G. d'Estaing capitalisés contrairement aux discours de F. Mitterrand et J. Chirac ; entretiens entre journalistes et hommes politiques.
2. **Lois européennes** : références remplacées par [REFERENCE] ; en-têtes, article final, lois de moins de 10 phrases supprimées ; numéros des articles, chapitres, titres et annexes ont été remplacés par la lettre "X".
3. **Ouvrage scientifique** : formules partiellement converties en texte ; références, citations remplacées par [REFERENCE].

Principales difficultés rencontrées 1/4

Difficultés mineures

- **Récupération** des discours politiques ; écriture d'un programme afin de récupérer un à un les fichiers sur le serveur.
- **Contrôle** qualité ; 30 Mo par personne ; vérifications au hasard dans les corpus sauf pour l'ouvrage scientifique.
- Fautes d'orthographes non corrigées notamment sur l'ouvrage scientifique.

Principales difficultés rencontrées 2/4

Difficulté moyenne

- **Remplacement** des formules dans l'ouvrage scientifique
- **Conversion** des formules les plus simples en texte, suppression de celles plus complexes
- Remplacement par [FORMULE] aussi envisageable.

Principales difficultés rencontrées 3/4

Difficulté majeure

- **Suppression** des en-tête et signature ; date terminant l'en-tête, signature de l'auteur du texte difficilement repérables.
- Biais d'apprentissage à supprimer pour ne pas rendre la tâche trop évidente.

Principales difficultés rencontrées 4/4

Difficulté majeure

- **Conversion** des fichiers depuis le format LaTeXde l'ouvrage scientifique ; pas de logiciel connu
- LaTeX→ HTML avec latex2html puis HTML → Texte avec Lynx

Conclusion sur la préparation

Conclusion

- La **conversion** des données pose un problème de complexité dû à la multiplicité des standards.
- La **segmentation** pose un problème d'anticipation dû à sa définition dépendante de son utilisation future.
- Le **contrôle** qualité pose un problème de temps dû à la taille toujours grandissante des corpus.

Plan

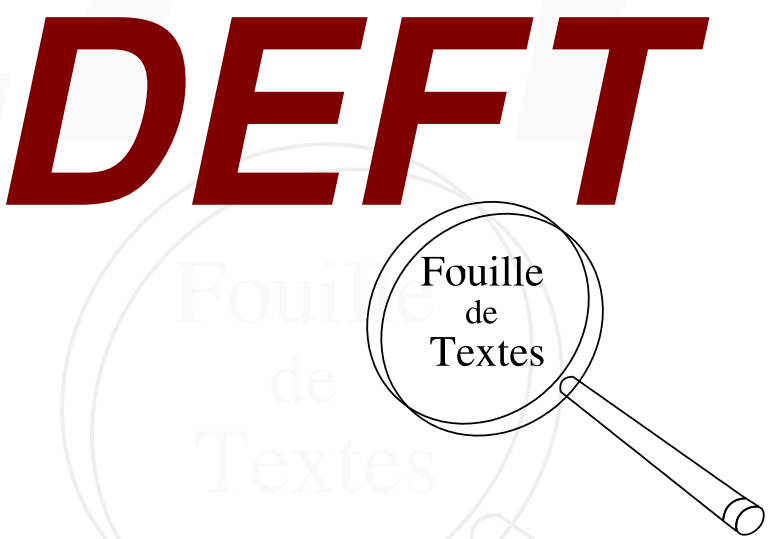
2 Naissance

3 Préparation

4 Organisation
- Tâches spécifiques au CO
- Calendriers

5 Évaluation et résultats

6 Annexes : résultats



Tâches spécifiques au Comité d'Organisation

Corpus

- Collecter le ou les corpus.
- Distribuer les rôles (dans le CO) pour la préparation du corpus.
- Créer les données d'apprentissage et de test.
- Définir le calendrier du défi (et le respecter).
- Assurer le plus rapidement possible les corrections d'éventuelles erreurs.

Tâches spécifiques au Comité d'Organisation

Relations avec les participants

- Assurer l'enregistrement de toutes les équipes.
- Collecter les éventuelles clauses de confidentialité.
- Relancer régulièrement les équipes inscrites mais n'ayant pas renvoyé un ou plusieurs documents.
- Assurer la bonne diffusion des informations (liste, forum).
- Récupérer dans les délais les soumissions des participants et les traiter rapidement pour détecter d'éventuels problèmes.
- Envoyer les résultats.
- Collecter les articles et les présentations (en temps voulu).

Tâches spécifiques au Comité d'Organisation

Relations avec le CP

- Constituer un CP en accord avec les présidents du CP.
- Faire valider par le CP le texte de l'appel du défi.
- Tenir régulièrement au courant le CP de l'évolution du défi (nombre d'équipes inscrites, nombre d'équipes participants réellement).
- Distribuer les articles pour relecture.
- Récupérer (dans les délais) les fiches de relecture ...

Calendrier de DEFT 2005

Principales dates pour l'année 2005

12 janvier	Diffusion du corpus d'apprentissage et ouverture des déclarations d'intention de participation
9 mars	Diffusion du 1^{er} corpus de test aux participants
11 mars	Soumission des résultats du 1^{er} corpus au comité, jusqu'à 20h Diffusion du $2^{ème}$ corpus de test aux participants, après 21h
15 mars	Soumission des résultats du $2^{ème}$ corpus au comité, jusqu'à 20h Diffusion du $3^{ème}$ corpus de test aux participants, après 21h
17 mars	Soumission des résultats du $3^{ème}$ corpus au comité, jusqu'à 20h
18 mars	Résultats diffusés aux participants
25 mars	Réception des articles
10 juin	Atelier de TALN'05 : 14h-17h30

Calendrier de DEFT 2006

Principales dates pour l'année 2006

début janvier	Diffusion de l'appel à communication et ouverture des déclarations d'intention de participation
13 février	Diffusion des corpus d'apprentissage
30 mai	Diffusion des corpus de test aux participants dès 9h
2 juin	Soumission des résultats au comité jusqu'à 17h
5 juin	Résultats diffusés aux participants
26 juin	Réception des articles
31 juillet	Notification aux auteurs
28 août	Réception de la version finale des articles
21-22 septembre	Atelier de la SDN'06

Conclusion sur l'organisation

Conclusion

- **Difficulté principale : Collecter et préparer les corpus ($\sim$ 3 mois)**
- Étapes de préparation des données laborieuses et demandant une bonne organisation et une juste répartition des tâches dans le CO.
- Coordination du défi (interface avec les participants et le CP) restreinte aux organisateurs ("chefs" du CO).
- Collecte des articles et redistribution au CP pour relecture : en général, le calendrier est assez tendu.

Plan

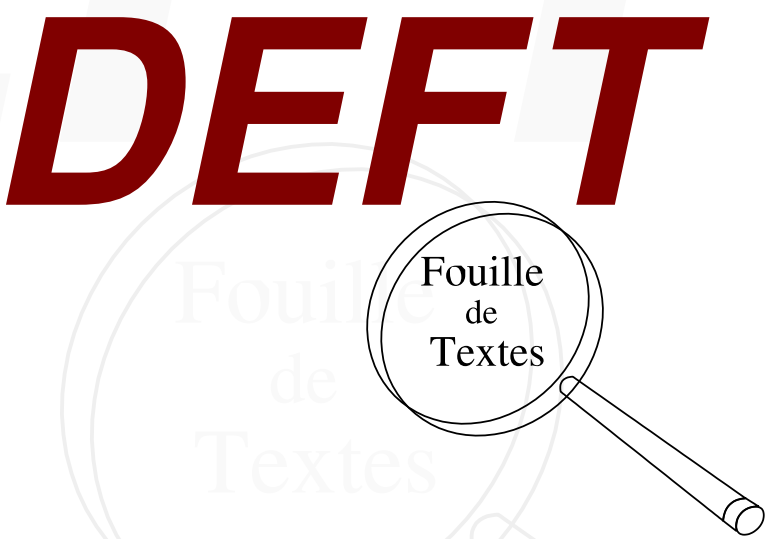
2 Naissance

3 Préparation

4 Organisation

5 Évaluation et résultats
- Critères d'évaluation
- Quelques résultats

6 Annexes : résultats



Critères d'évaluation utilisés dans DEFT

Matrice de confusion

Prédit \ *Réel*	Pos	Neg
Pos	TP	FP
Neg	FN	TN

- TP : True Positive
- FP : False Positive
- FN : False Negative
- TN : True Negative

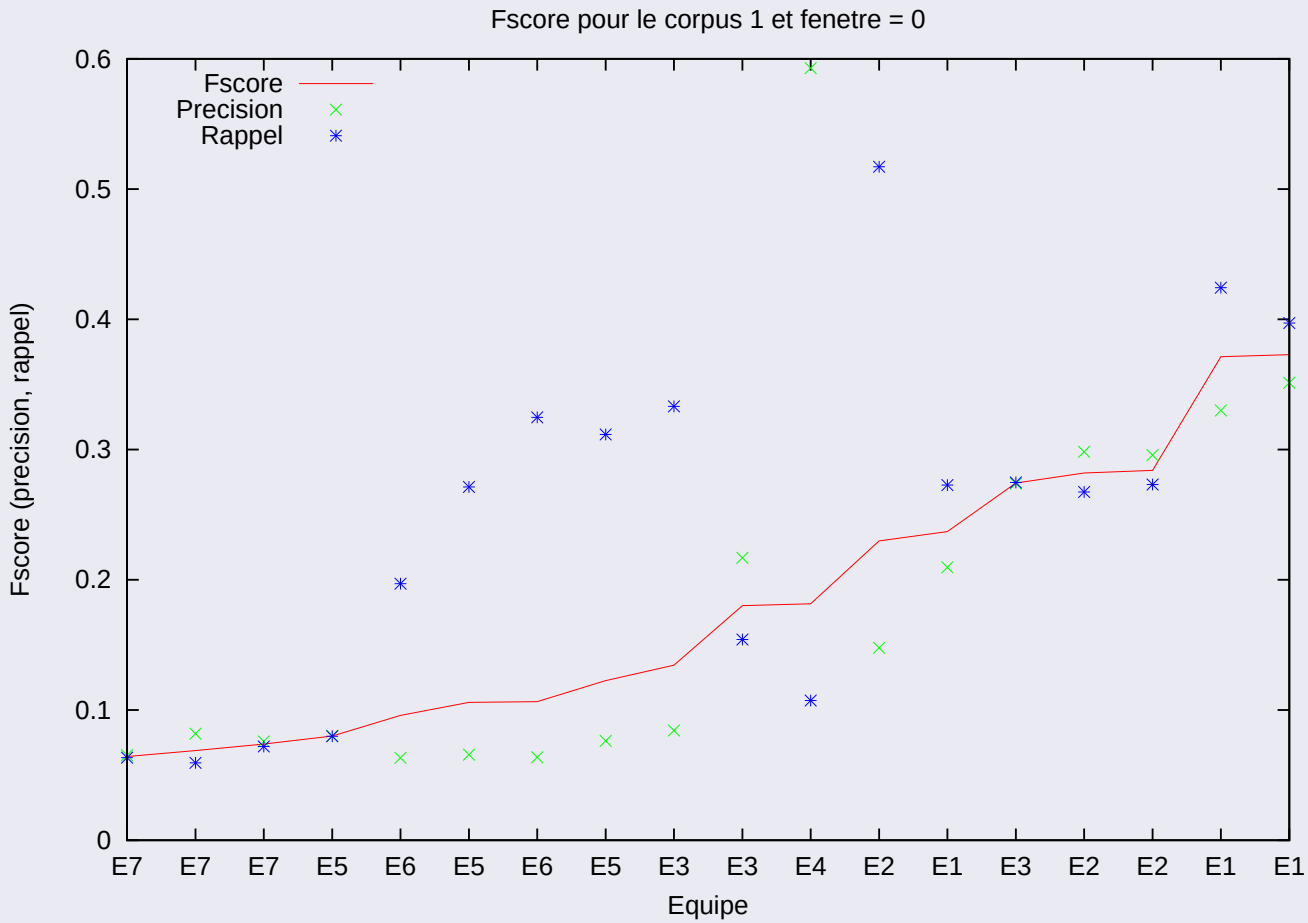
Précision, Rappel, F_{score}

- $Precision = \frac{TP}{TP+FP}$
- $Rappel = \frac{TP}{TP+FN}$
- $F_{score} = \frac{(\beta^2+1) \times Precision \times Rappel}{\beta^2 \times Precision + Rappel}$

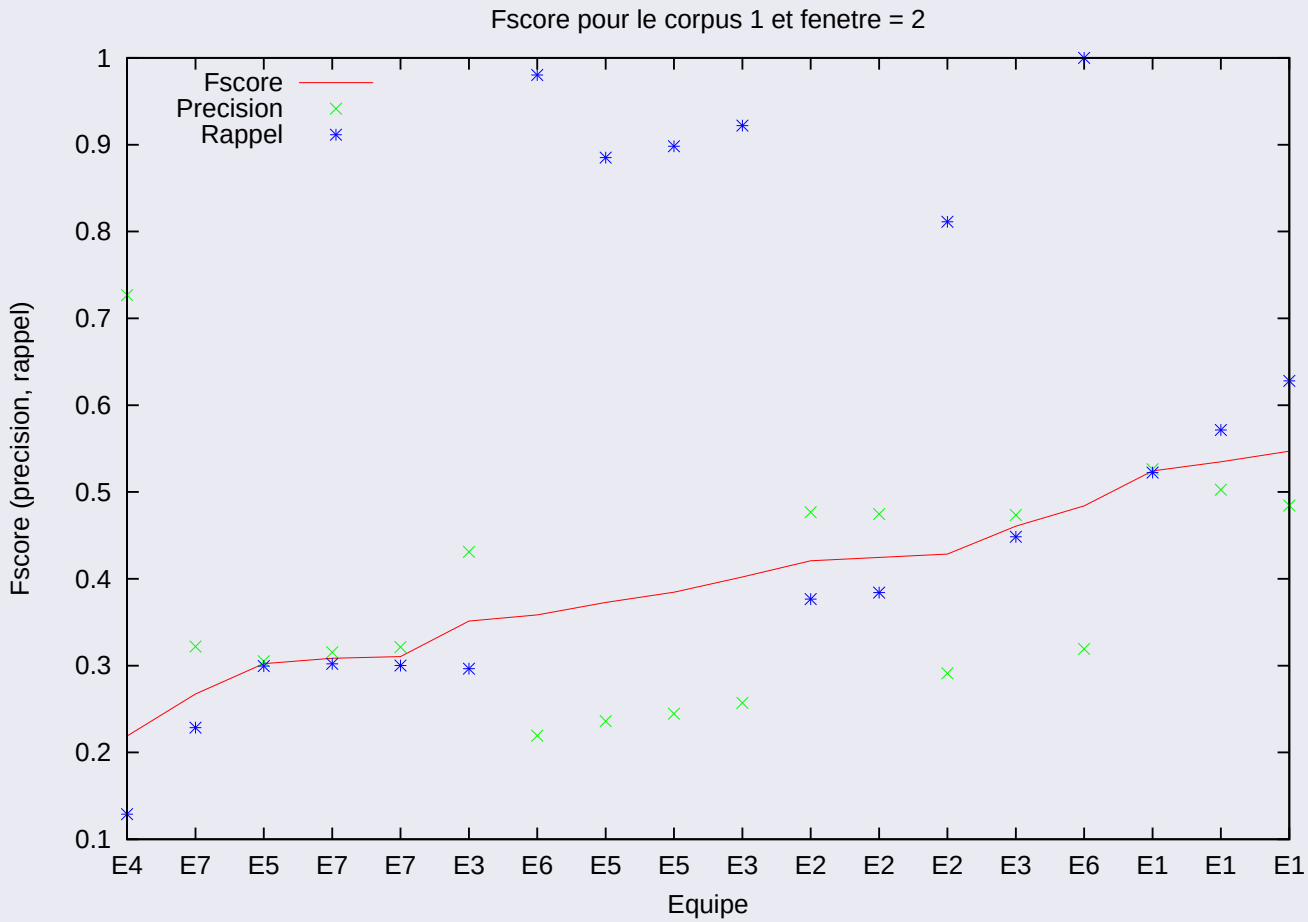
Front de Pareto, F_{score} souple

- F_{score} souple : F_{score} avec tolérance de $\pm$ 1 ou 2 autour de la solution
- Front de Pareto : ensemble des approches qui sont telles qu'aucune autre approche ne présente de meilleurs résultats pour tous les critères étudiés

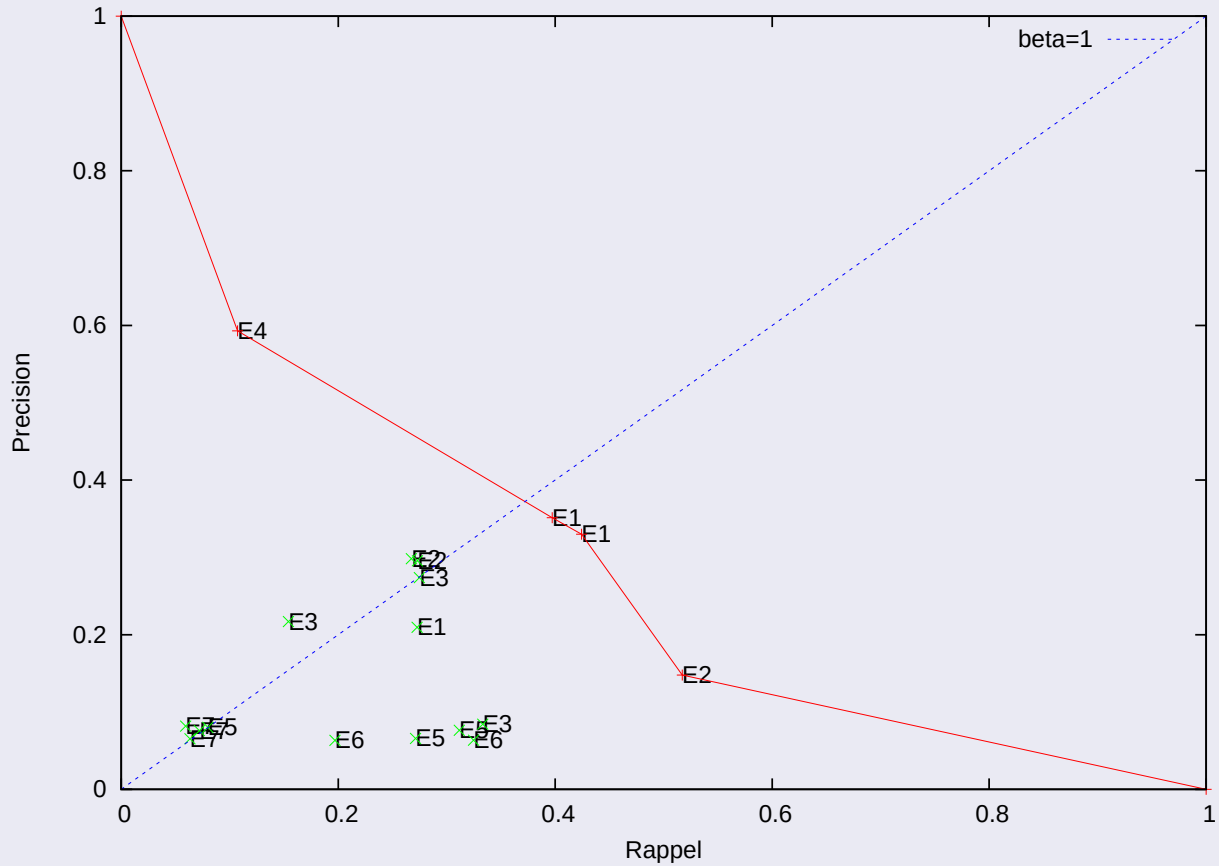
DEFT 2006, Corpus "Discours politique", F_{score} strict



DEFT 2006, Corpus "Discours Politique", F_{score} souple, fenêtre = 2



DEFT 2006, Front de Pareto pour les "Discours Politiques"



Plan

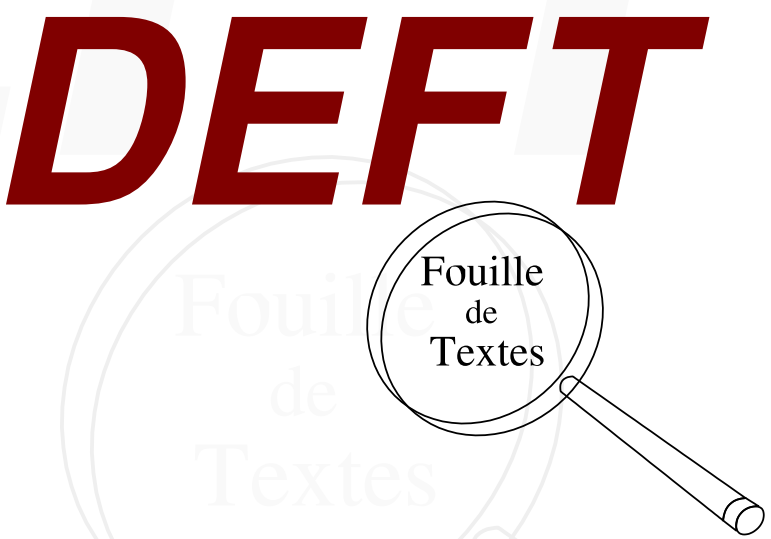
2 Naissance

3 Préparation

4 Organisation

5 Évaluation et résultats

6 Annexes : résultats



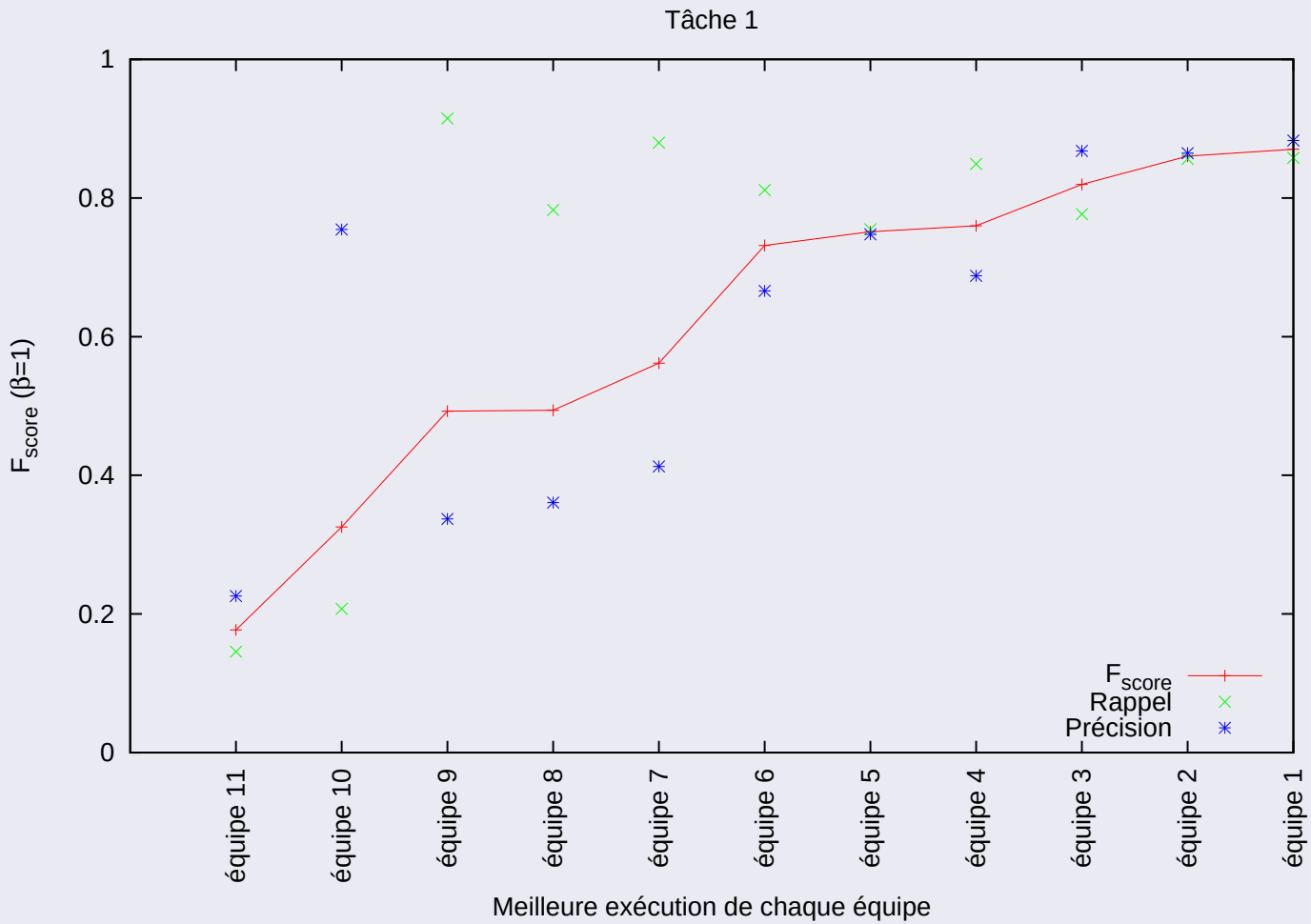
Critères d'évaluation

DEFT 2005

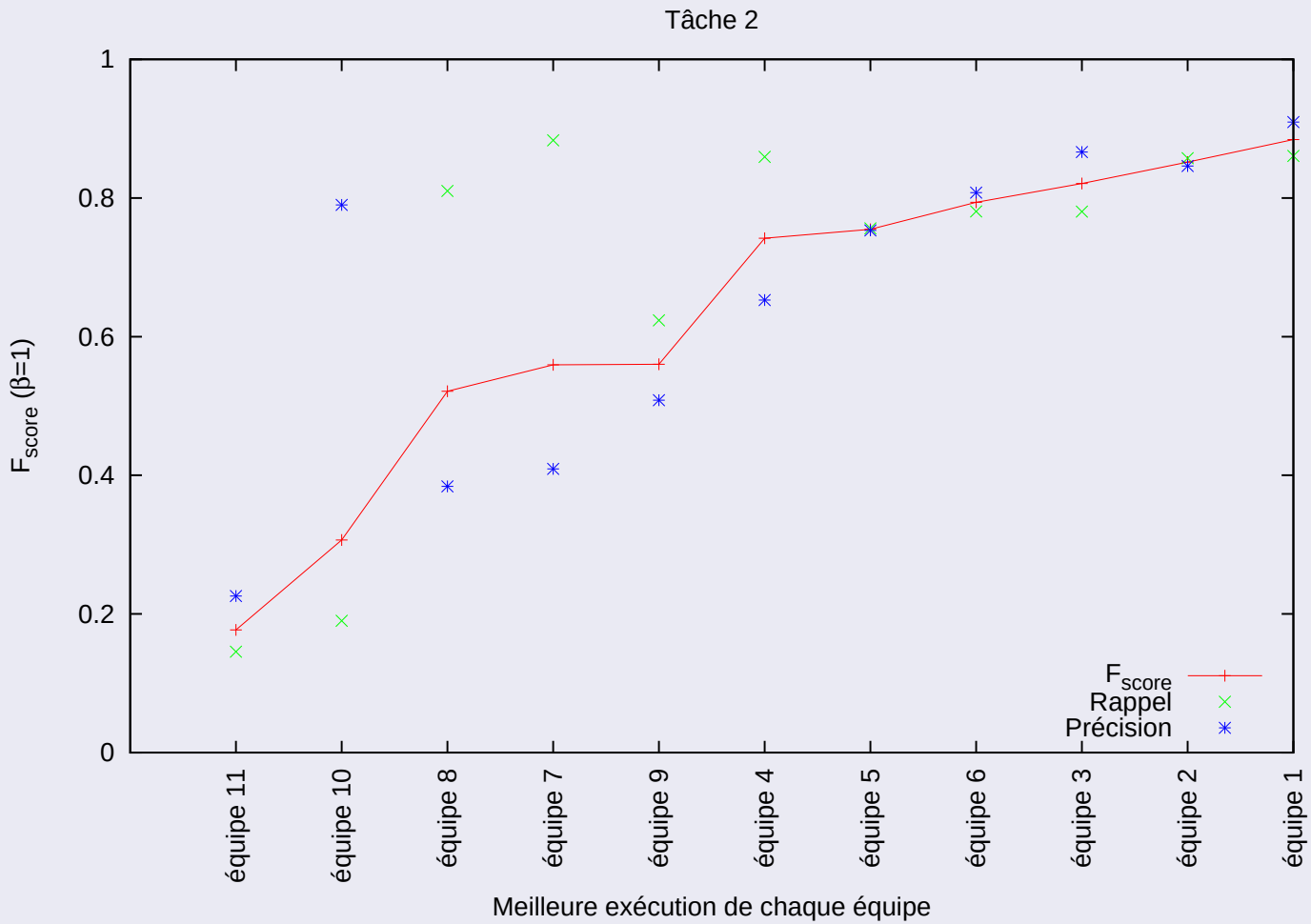
- Précision : pourcentage de phrases correctement associées à F. Mitterrand dans le fichier résultat parmi les phrases soumises.
- Rappel : pourcentage de phrases correctement associées à F. Mitterrand dans le fichier résultat parmi les phrases de F. Mitterrand réellement introduites dans le corpus.
-

$$F_{score}(\beta = 1) = \frac{(\beta^2 + 1) \times Précision \times Rappel}{\beta^2 \times Précision + Rappel} \quad (1)$$

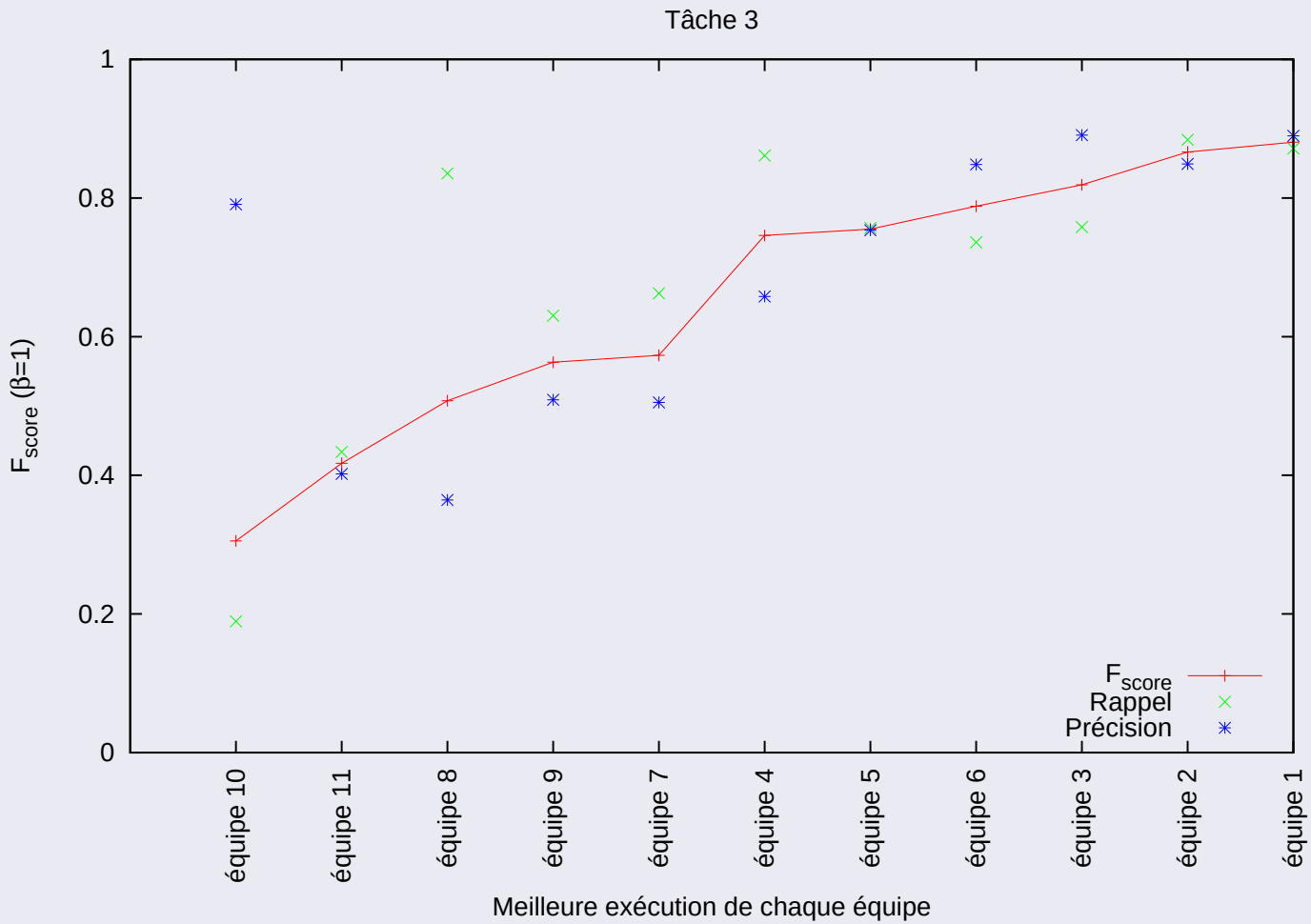
F_{scores} ($\beta = 1$) pour les meilleures exécutions - Tâche 1.



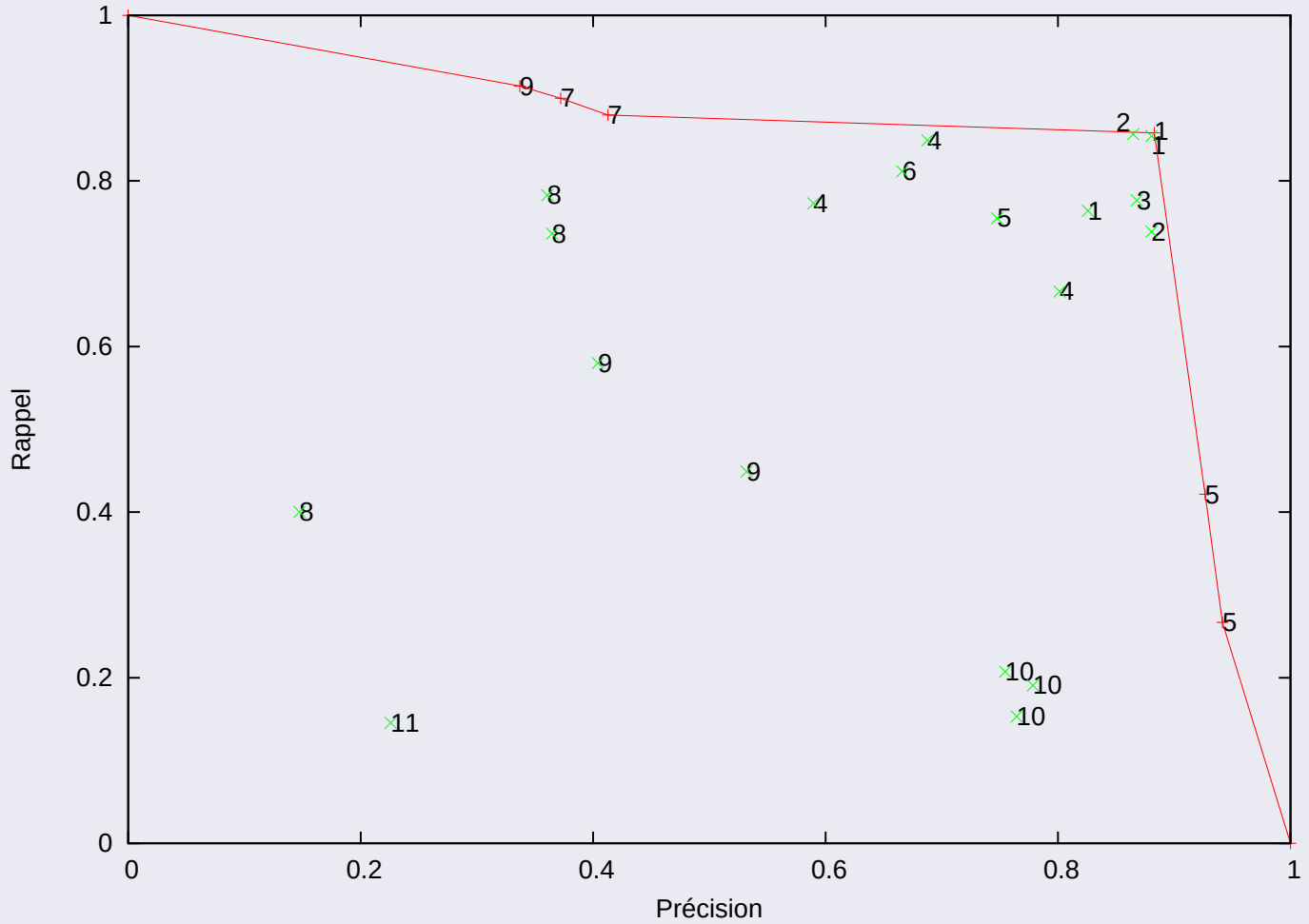
F_{scores} ($\beta = 1$) pour les meilleures exécutions - Tâche 2.



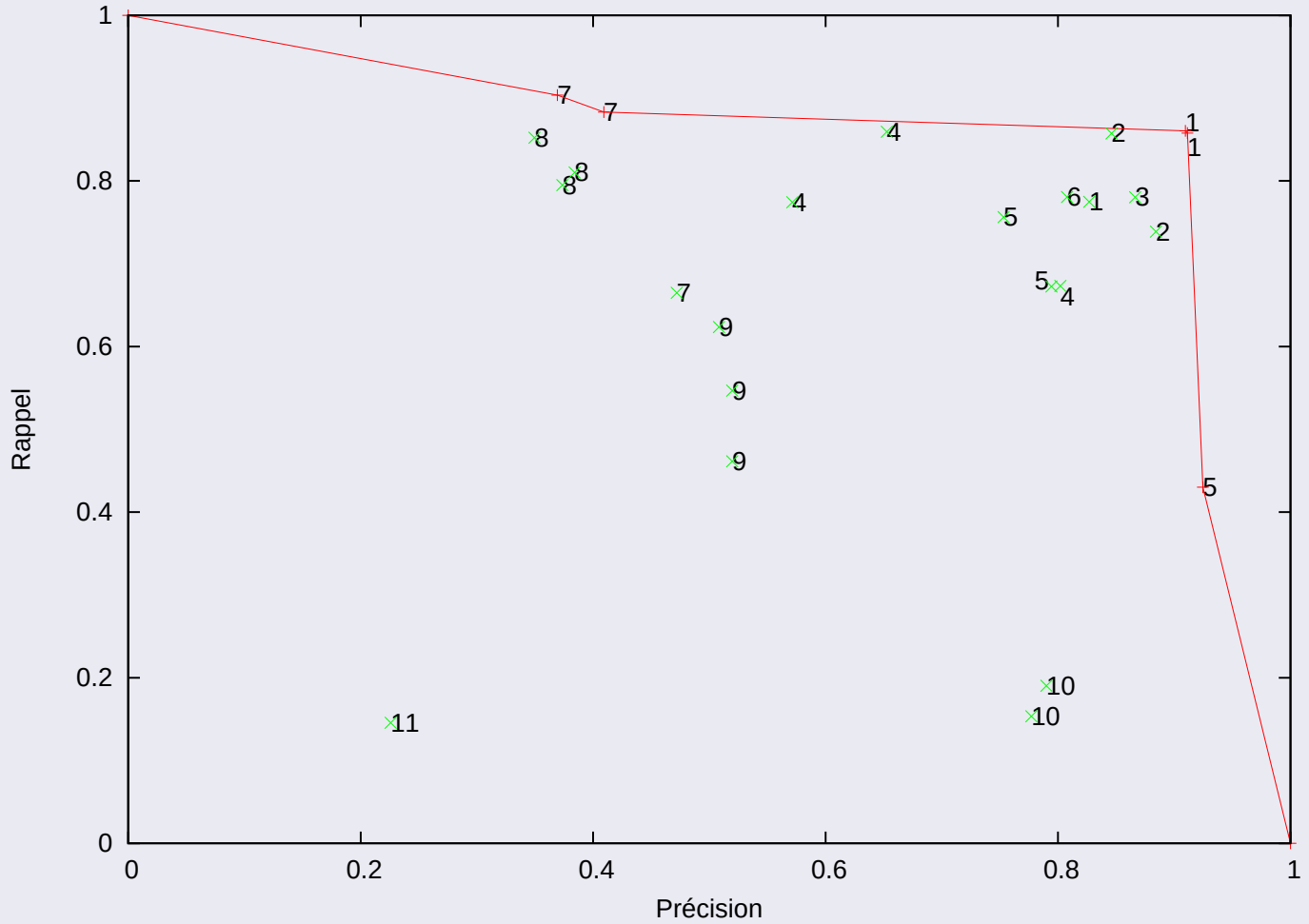
F_{scores} ($\beta = 1$) pour les meilleures exécutions - Tâche 3.



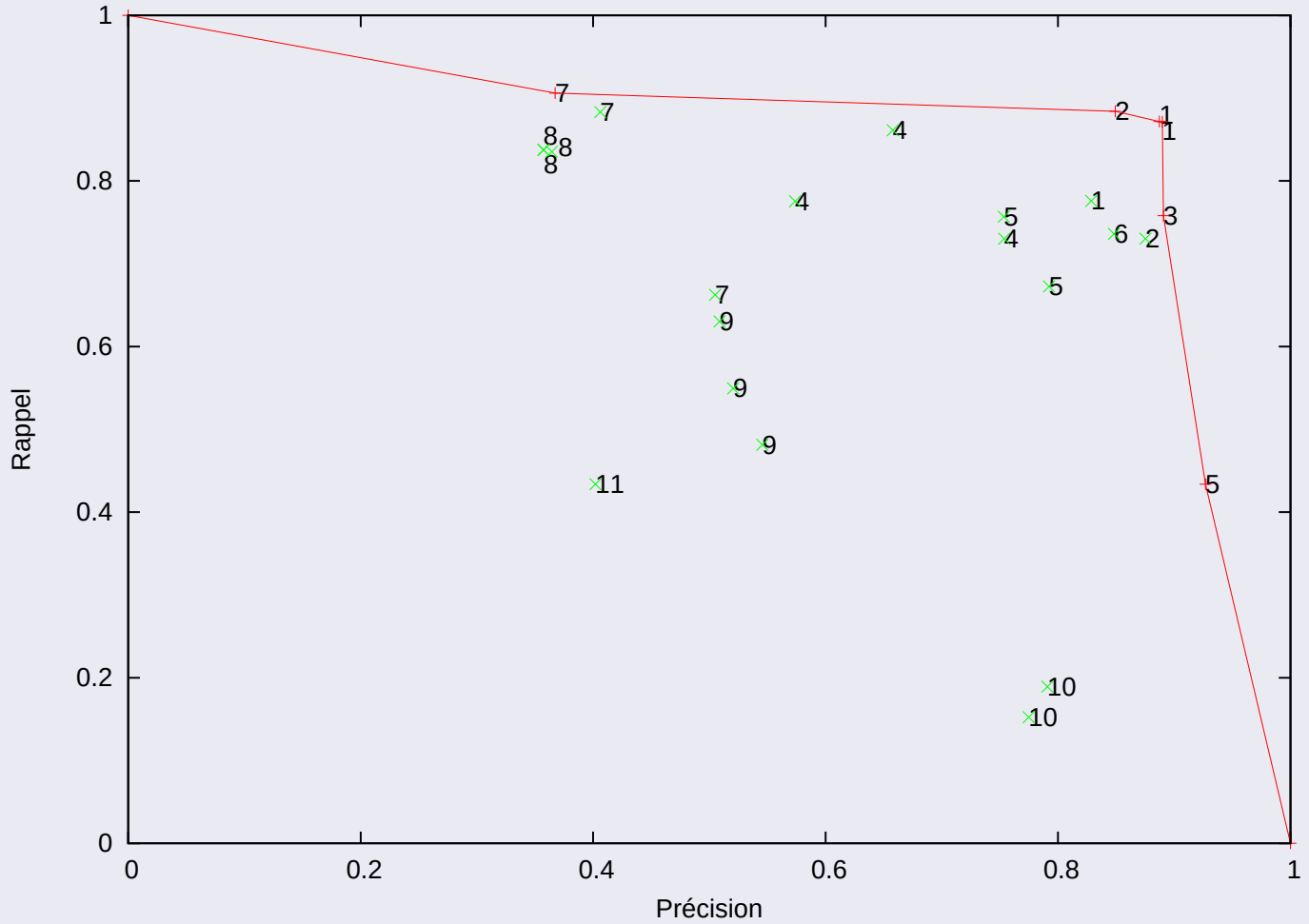
Front de Pareto pour la tâche 1.



Front de Pareto pour la tâche 2.



Front de Pareto pour la tâche 3.



Résultats

Meilleurs F_{score} des différentes équipes pour chaque tâche.

	tâche 1	**tâche 2**	**tâche 3**
équipe 1	0.870 *(1)*	0.884 *(1)*	0.880 *(1)*
équipe 2	0.860 *(2)*	0.852 *(2)*	0.866 *(2)*
équipe 3	0.820 *(3)*	0.821 *(3)*	0.819 *(3)*
équipe 4	0.760 *(4)*	0.742 *(6)*	0.746 *(6)*
équipe 5	0.751 *(5)*	0.755 *(5)*	0.755 *(5)*
équipe 6	0.732 *(6)*	0.794 *(4)*	0.788 *(4)*
équipe 7	0.562 *(7)*	0.559 *(8)*	0.573 *(7)*
équipe 8	0.494 *(8)*	0.521 *(9)*	0.507 *(9)*
équipe 9	0.493 *(9)*	0.560 *(7)*	0.563 *(8)*
équipe 10	0.325 *(10)*	0.307 *(10)*	0.305 *(11)*
équipe 11	0.177 *(11)*	0.177 *(11)*	0.417 *(10)*

Plan

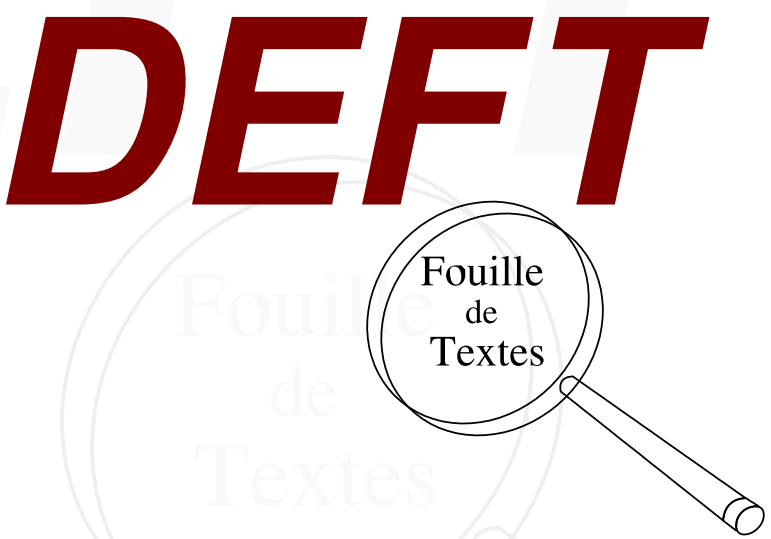
2 Naissance

3 Préparation

4 Organisation

5 Évaluation et résultats

6 Annexes : résultats



Précision, Rappel, F_{score}

Utilisation des "mesures classiques" d'évaluation des soumissions effectuées

Précision

$$Précision = \frac{\#\text{ segments thématiques correctement identifiés}}{\#\text{ segments thématiques extraits}}$$

Rappel

$$Rappel = \frac{\#\text{ segments thématiques correctement identifiés}}{\#\text{ segments thématiques total}}$$

F_{score}

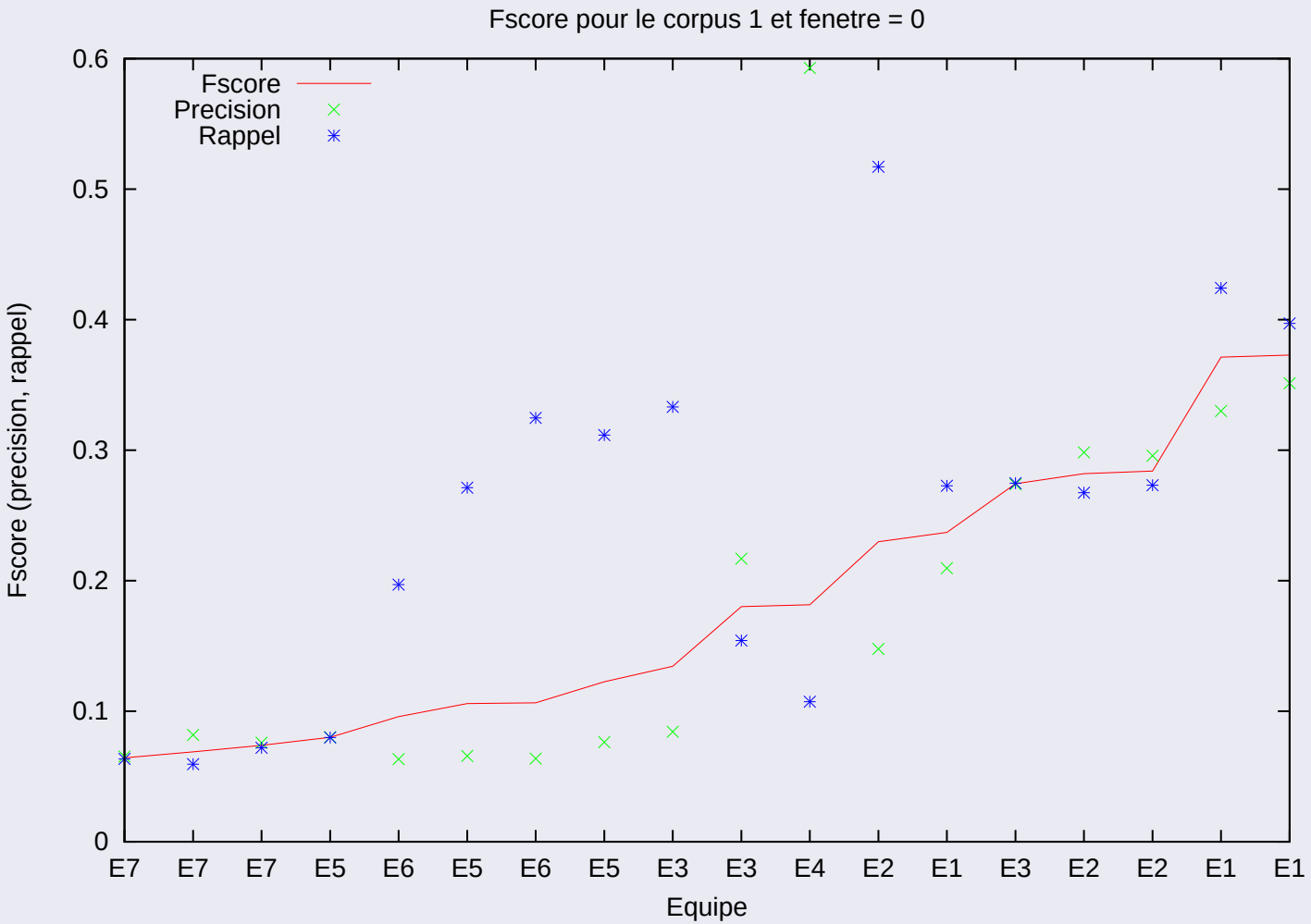
F_{score}

$$F_{score}(\beta) = \frac{(\beta^2 + 1) \times Précision \times Rappel}{\beta^2 \times Précision + Rappel}$$

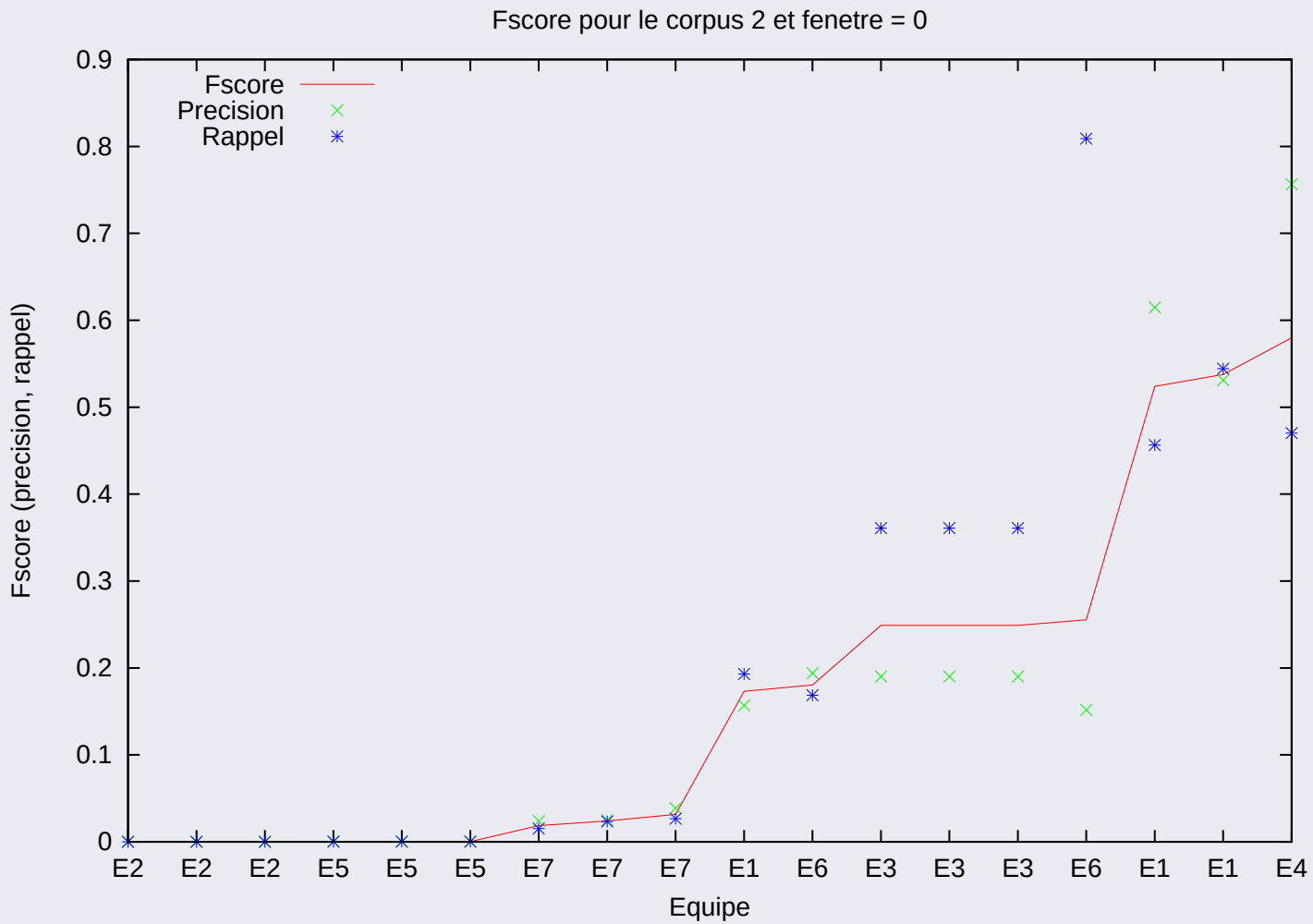
Dans le cadre de DEFT, le paramètre β a été fixé à 1, soit :

$$F_{score} = \frac{2 \times (\#\text{ segments thématiques correctement identifiés})}{\#\text{ segments thématiques extraits} + \#\text{ segments thématiques total}}$$

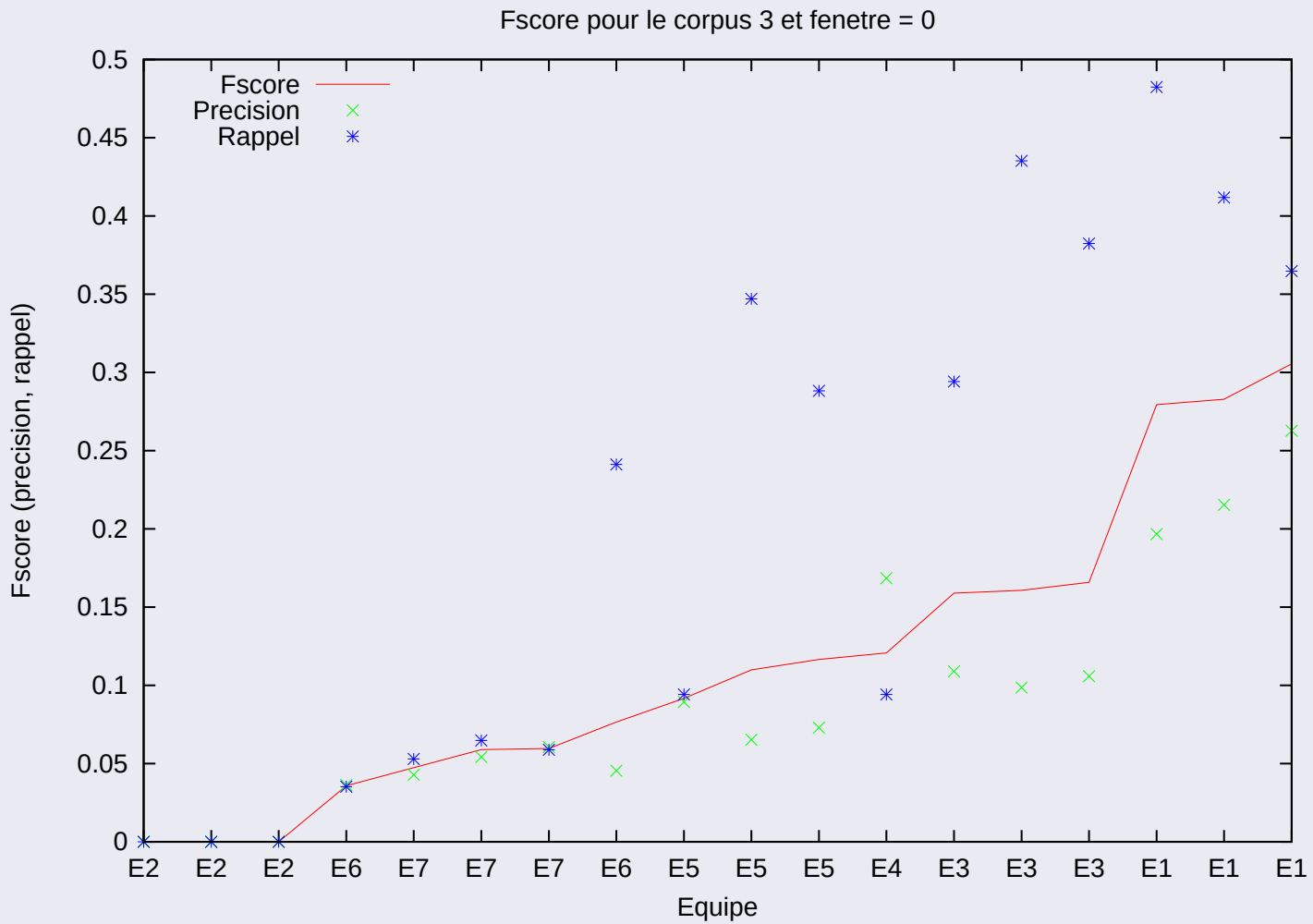
Corpus "Discours politique", F_{score} strict



Corpus "Lois", F_{score} strict



Corpus "Ouvrage scientifique", F_{score} strict

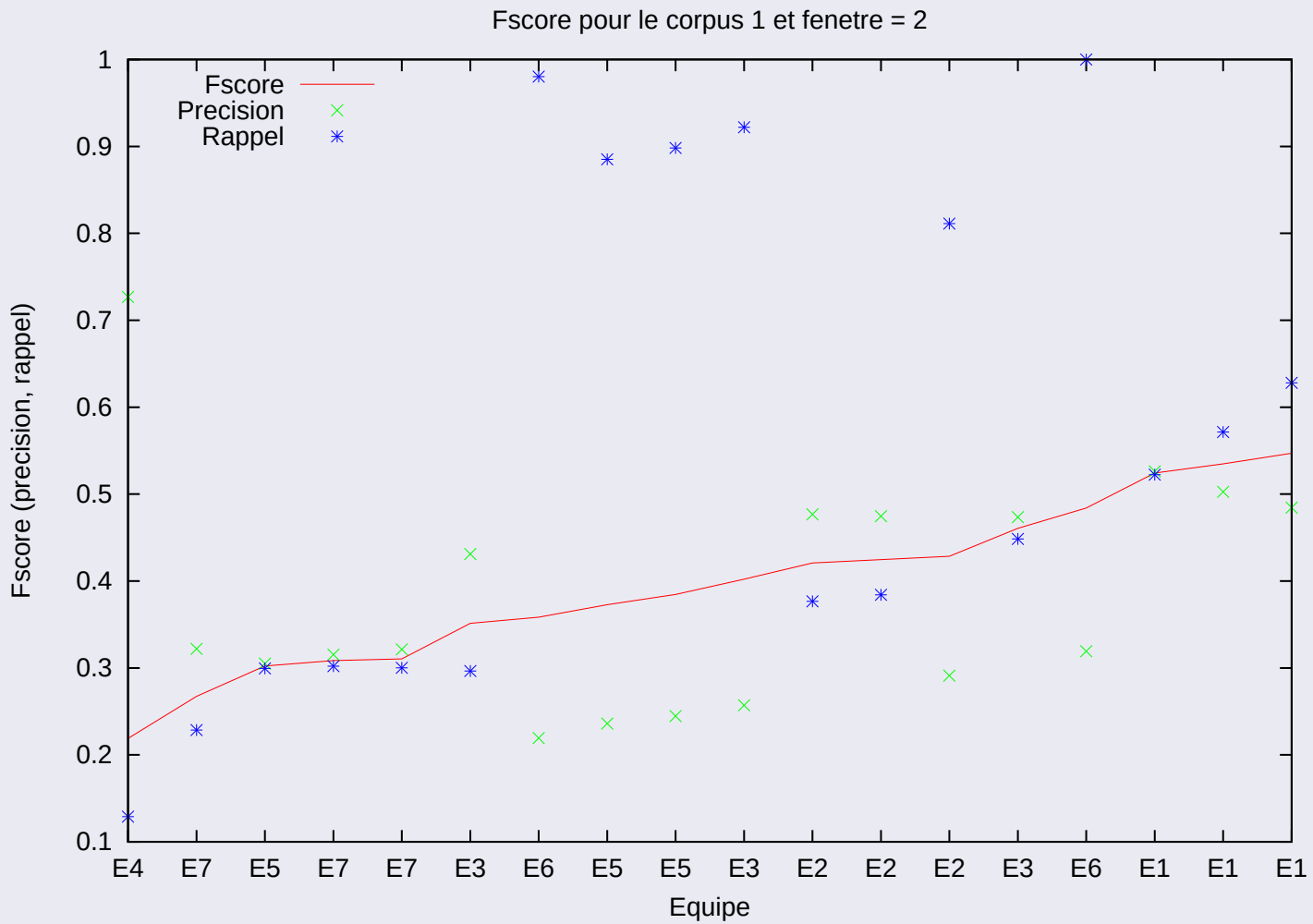


Évaluation souple

F_{score} souple

- Compte tenu de la difficulté de la tâche, il est intéressant d'évaluer les résultats obtenus en tolérant une marge d'erreur autour de la solution exacte
- Une tolérance de $\pm$ une(deux) phrase(s) autour de la phrase marquant le début du segment a été introduite.
- Les mesures de précision, rappel et F_{score} ont été modifiées en conséquence

Corpus "Discours Politique", F_{score} souple, fenêtre = 2



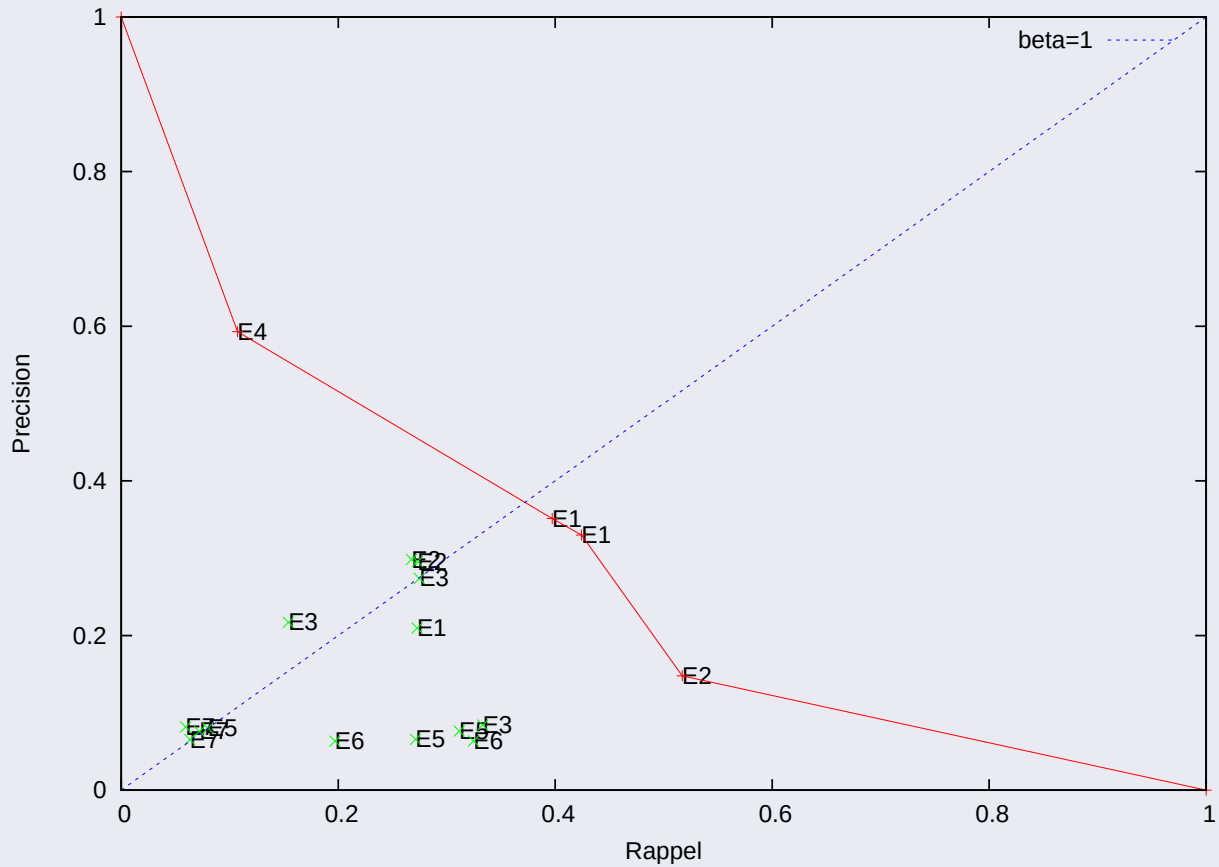
Front de Pareto 1/2

Définition

- Le front de Pareto est défini par l'ensemble des approches qui sont telles qu'aucune autre approche ne présente de meilleurs résultats pour tous les critères étudiés (ici précision et rappel).
- Les approches qui ne sont pas sur le front de Pareto sont dites "dominées".

Front de Pareto 2/2

Front de Pareto pour les "Discours Politiques"



Tri des soumissions et désignation du vainqueur

Principe

- Les soumissions sont considérées comme indépendantes.
- Le F_{score} strict est utilisé pour trier les soumissions.
- Pour chaque équipe, la soumission ayant le rang global (somme des rangs pour chaque corpus) le plus faible est retenue.
- L'équipe ayant le rang le plus faible est désignée comme vainqueur.

Palmarès - 18 soumissions

	Discours Politique	Lois	Ouvrage Scientifique
Hurault-Plantet et al. (E1) LIMSI	1	3	2
Khalis et al. (E3) CLIPS - IMAG	5	5	4
Da Sylva et al. (E4) GRDS	8	1	7
Labadié et Chauché (E6) LIRMM	12	4	11
Wildocher et al. (E5) GREYC CNRS UMR 6072	11	13	8
Trinh et Galinari (E7) LIP6	16	10	13
Lelu et al. (E2) LASELDI	3	–	–

Palmarès avec F_{score} souple - 18 soumissions

	Discours		**Lois**		**Ouvrage**	
	Fenêtre		*Fenêtre*		*Fenêtre*	
	1	2	1	2	1	2
Hurault-Plantet et al. (E1) LIMSI	1	2	2	2	2	2
Khalis et al. (E3) CLIPS - IMAG	4	5	6	5	5	5
Da Sylva et al. (E4) GRDS	14	18	1	1	10	11
Labadié et Chauché (E6) LIRMM	12	12	5	8	9	8
Wildocher et al. (E5) GREYC CNRS UMR 6072	10	10	10	10	8	7
Trinh et Galinari (E7) LIP6	16	15	14	13	12	10
Lelu et al. (E2) LASELDI	5	6	–	–	–	–

Conclusion sur les résultats - 2005

Comparaison des traitements effectués par les équipes.

Numéro d'équipe	**1**	**2**	3	**4**	**5**	6	**7-9**	8	**10**
Types de prétraitements									
Suppression de mots		✓		✓				✓	✓
N-lettres			7						
N-mots	1-2		4	1-3			4		
Méthodes de classification									
Chaînes de Markov	✓	✓		✓		✓	✓		
Viterbi	✓	✓		✓			✓		
Bayes	✓		✓	✓				✓	
SVM				✓		✓			
Apports linguistiques									
Vecteurs de termes								✓	✓
Étiquetage grammatical	✓					✓	✓	✓	✓
Relations syntaxiques					✓				✓
Entités nommées	✓					✓			

Questions

Questions . . .