DEFT'07 – 3ème Défi Fouille de Texte
3 juillet 2007, Grenoble

Présentation du défi et résultats

LIMSI-CNRS & LRI/Université Paris Sud

Précédentes éditions de DEFT

- **2005 :**
  - **Identification de locuteurs dans des discours politiques (discours de F. Mitterrand et J. Chirac) ;**
  - **Tâche de profilage et de segmentation stylistique.**
- **2006 :**
  - **Segmentation thématique de textes politiques (discours), scientifiques (livres) et juridiques (lois européennes) ;**
  - **Tâche de cohésion stylistique (vocabulaire, style du texte, marqueurs) et thématique (recherche des discontinuités dans les thèmes).**

Édition 2007 de DEFT

- **Objectif : attribuer automatiquement des valeurs d'opinion à des textes présentant un avis argumenté – positif, négatif ou éventuellement mitigé – sur un sujet donné.**
- **Tâche de classification.**
- **Domaine applicatif :**
  - **Entreprises : analyse automatique de données, aide à la prise de décision ;**
  - **Consommateurs : tri des évaluations de produits par les consommateurs sur l'Internet, production d'un jugement conclusif.**

Présentation des corpus

- **« À voir, à lire » :**
  - **Environ 3 000 critiques (7,6 Mo) de livres, de films et de spectacles provenant du site www.avoir-alire.com**
- **Jeux vidéos :**
  - **Environ 4 000 critiques (28,3 Mo) de jeux vidéos provenant du site www.jeuxvideos.com**
- **Relectures d'articles scientifiques :**
  - **Environ 1 000 relectures d'articles (2,4 Mo) issues des campagnes JADT, RFIA et TALN.**
- **Débats parlementaires :**
  - **Environ 28 800 interventions de Députés (38,1 Mo).**

Préparation des données

- **Réencodage des accents en iso-latin-1 avec élimination des accents hors de cette norme (*Tōkyō* → *Tokyo*) ;**
- **Homogénéisation des fins de ligne (\r\n → \n) ;**
- **Conversion des documents au format XML ;**
- **Quelques traitements spécifiques :**
  - **Conversion préalable des documents Word, Excel et LaTeX en texte brut pour le corpus des relectures ;**
  - **Anonymisation des données pour les corpus des relectures et des débats parlementaires ;**
  - **Élimination des interventions de moins de 300 caractères pour le corpus des débats parlementaires.**

Évaluations manuelles

- **Évaluations manuelles d'extraits de chaque corpus avec essais de différentes échelles pour les classes d'opinion.**
- **Confrontation des résultats par le coefficient Kappa.**

Juge	Réf.	1	2
Réf.		0,17	0,12
1	0,17		0,03
2	0,12	0,03	

Juge	Réf.	1	2
Réf.		0,74	0,79
1	0,74		0,74
2	0,79	0,74	

- **Échelle de 0 à 20 (à gauche), de 0 à 2 (à droite), sur le corpus des jeux vidéos.**
- **Meilleur accord sur l'échelle restreinte.**

Classes d'opinion par corpus

- **« À voir, à lire » :**
  - **2 (favorable), 1 (neutre) et 0 (défavorable).**
- **Jeux vidéos :**
  - **2 (avis positif), 1 (avis moyen) et 0 (avis négatif).**
- **Relectures d'articles scientifiques :**
  - **2 (article accepté en l'état ou après modifications mineures) ;**
  - **1 (article accepté après modifications majeures) ;**
  - **0 (article rejeté).**
- **Débats parlementaires :**
  - **1 (vote favorable à la loi en examen) et 0 (vote défavorable).**

Indices de confiance

- **L'indice de confiance est la probabilité pour un document d'appartenir à une classe donnée.**
- **Utilisation des indices de confiance pour permettre l'attribution de plusieurs classes d'opinion à un même document.**
- **Introduit une pondération de la confiance et du rappel.**

Evaluation des résultats

- **Les mesures**
  - **F-score**
  - **F-score pondéré par l'indice de confiance**
- **Le classement**
  - **Par soumission, en calculant un score global pour chaque soumission**

F-score

- **Formule générale**

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

- **Moyennes globales de la précision et du rappel pour n classes**
  - **Micro-moyenne**

$$\text{Précision} = \sum_{i=1}^{n} TP_i \Big/ \sum_{i=1}^{n} (TP_i + FP_i) \qquad Rappel = \sum_{i=1}^{n} TP_i \Big/ \sum_{i=1}^{n} (TP_i + FN_i)$$

  - **Macro-moyenne**

$$\text{Précision} = \sum_{i=1}^{n} (TP_i / (TP_i + FP_i)) \Big/ n \qquad Rappel = \sum_{i=1}^{n} (TP_i / (TP_i + FN_i)) \Big/ n$$

  - **Avec :**
    - FP_i : nombre de documents faussement attribués à la classe *i*
    - TP_i : nombre de documents correctement attribués à la classe *i*
    - FN_i : nombre de documents appartenant à la classe *i* et non retrouvés par le système
    - *n* : nombre de classes

F-score pondéré par l'indice de confiance

- **Utilisation du *F-score* avec *β*=1 et la macro-moyenne de la précision et du rappel sur les *n* classes**
- **Précision et rappel pondérés par l'indice de confiance**

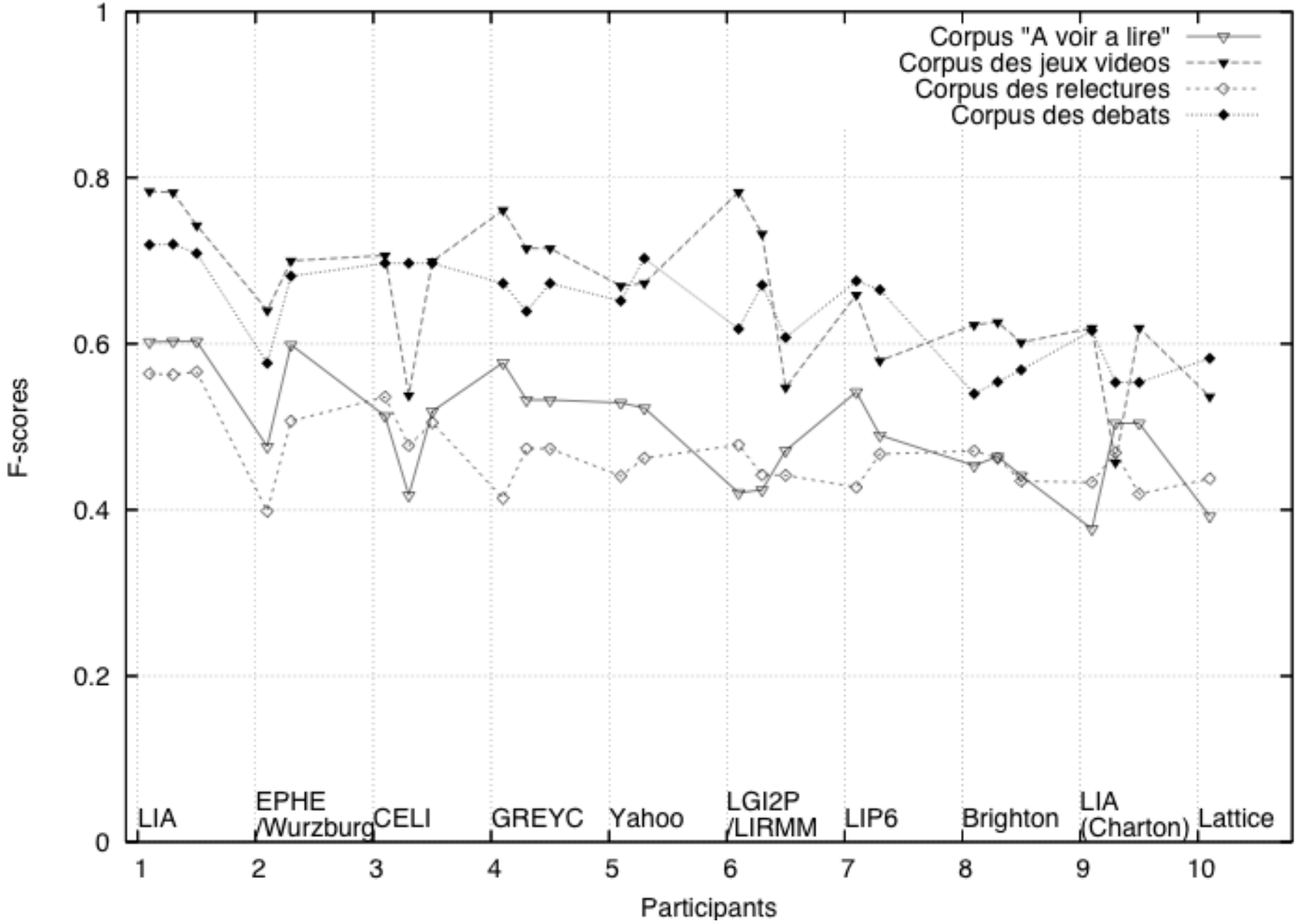
$$\text{Précision}_i = \sum_{d=1}^{Nac_i} indice_d^i \Big/ \sum_{d=1}^{Na_i} indice_d^i \qquad Rappel_i = \sum_{d=1}^{Nac_i} indice_d^i \Big/ N_i$$

- **Avec :**
  - ***Nac*$_i$: nombre de documents appartenant effectivement à la classe *i* et auquels le système a attribué un indice de confiance non nul pour cette classe**
  - ***Na*$_i$: nombre de documents auxquels le système a attribué un indice de confiance non nul pour la classe *i***
  - ***N*$_i$: nombre de documents appartenant à la classe *i***

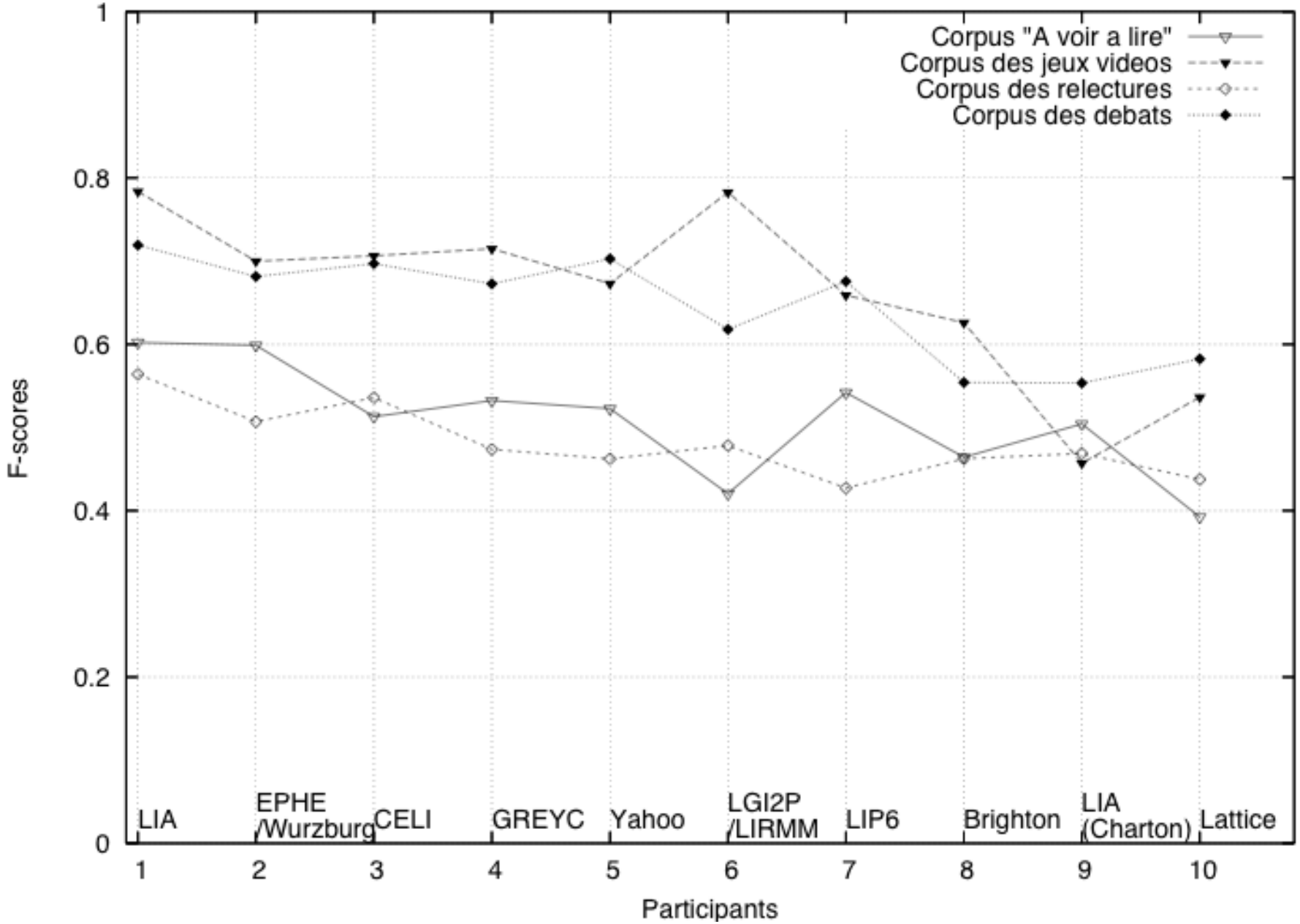
Classement des soumissions

- **Chaque soumission est classée**
  - **Pour chaque *corpus*, chaque soumission est classée dans l'ordre du *F-score* décroissant, et obtient donc un *rang(corpus, soumission)***
  - **Pour chaque *soumission*, la somme des *rangs* obtenus pour chaque *corpus* nous donne un *score-rang(soumission)***
  - **Les *soumissions* sont ensuite classées dans l'ordre croissant du *score-rang***
- **L'équipe gagnante est celle qui a la soumission la mieux classée**

F-scores stricts par corpus (ensemble des soumissions)



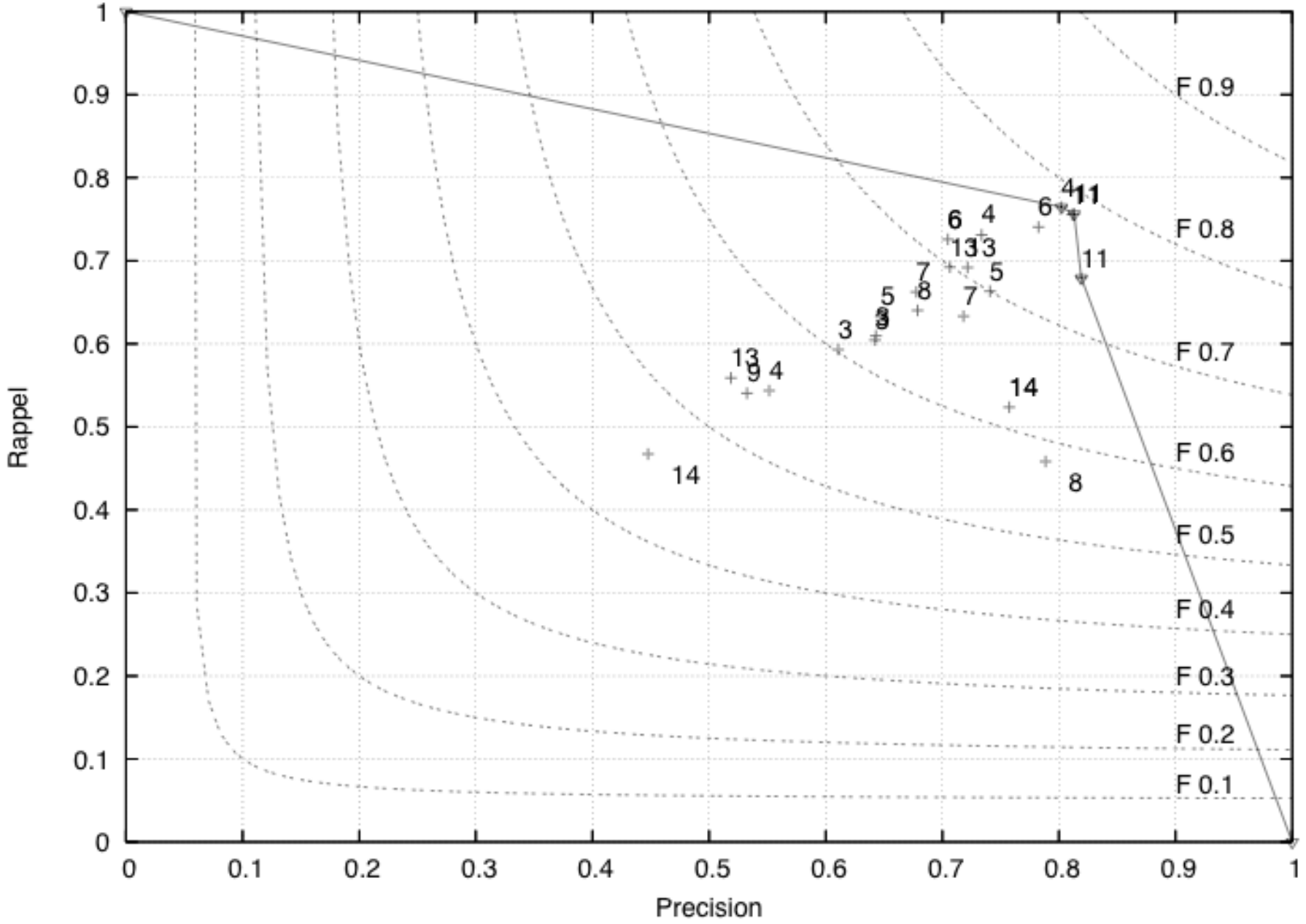
F-scores stricts par corpus (meilleures soumissions)



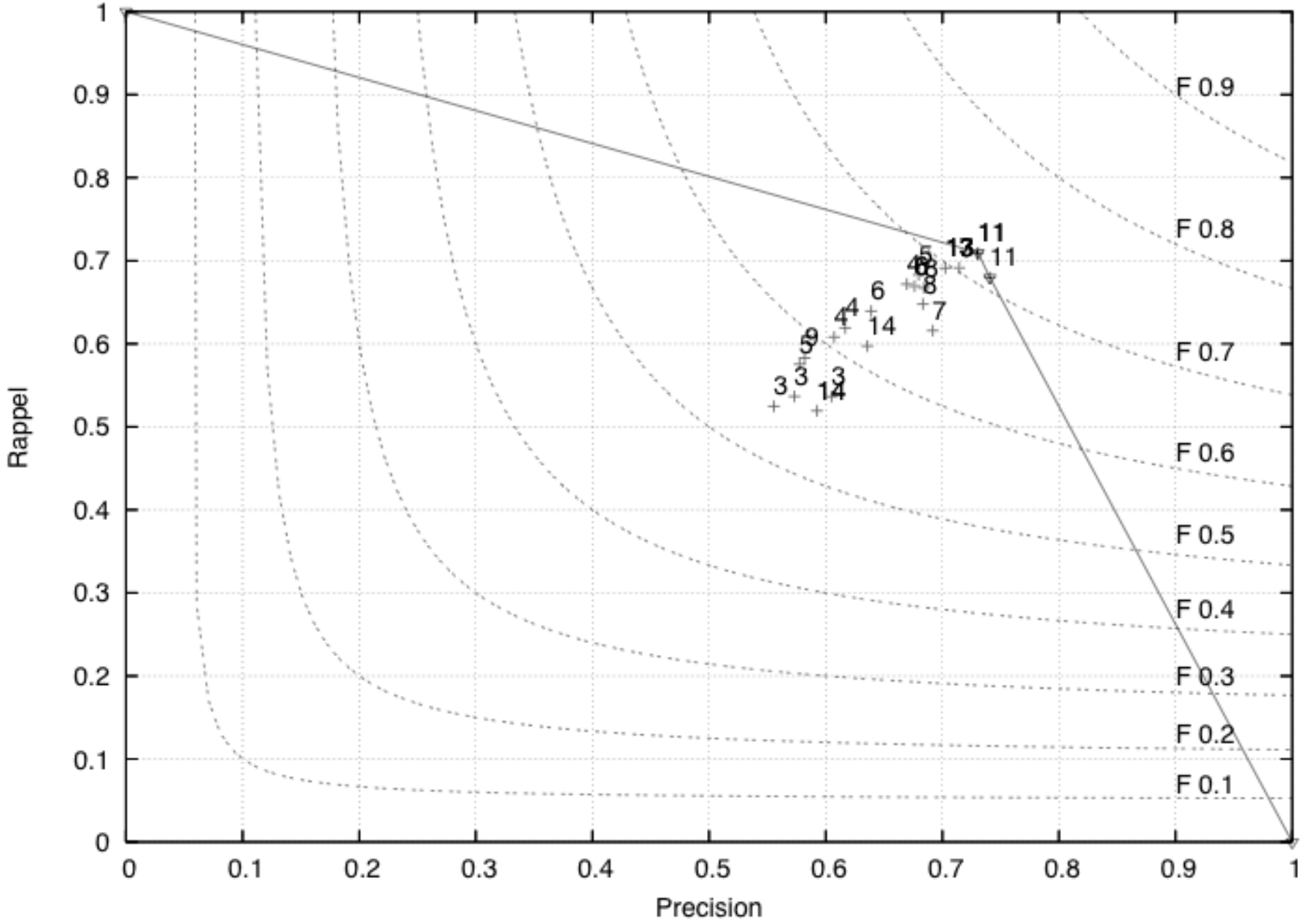
F-scores stricts par corpus (meilleures soumissions)

Équipe	À voir, à lire	Jeux vidéos	Relectures	Débats
LIA	0,602	0,784	0,564	0,719
EPHE	0,599	0,699	0,507	0,681
CELI	0,513	0,706	0,536	0,697
GREYC	0,532	0,715	0,474	0,673
Yahoo!	0,523	0,673	0,462	0,703
LGI2P	0,421	0,783	0,478	0,618
LIP6	0,542	0,659	0,427	0,676
Brighton	0,464	0,626	0,463	0,554
LIA Charton	0,504	0,457	0,469	0,553
Lattice	0,392	0,536	0,437	0,582

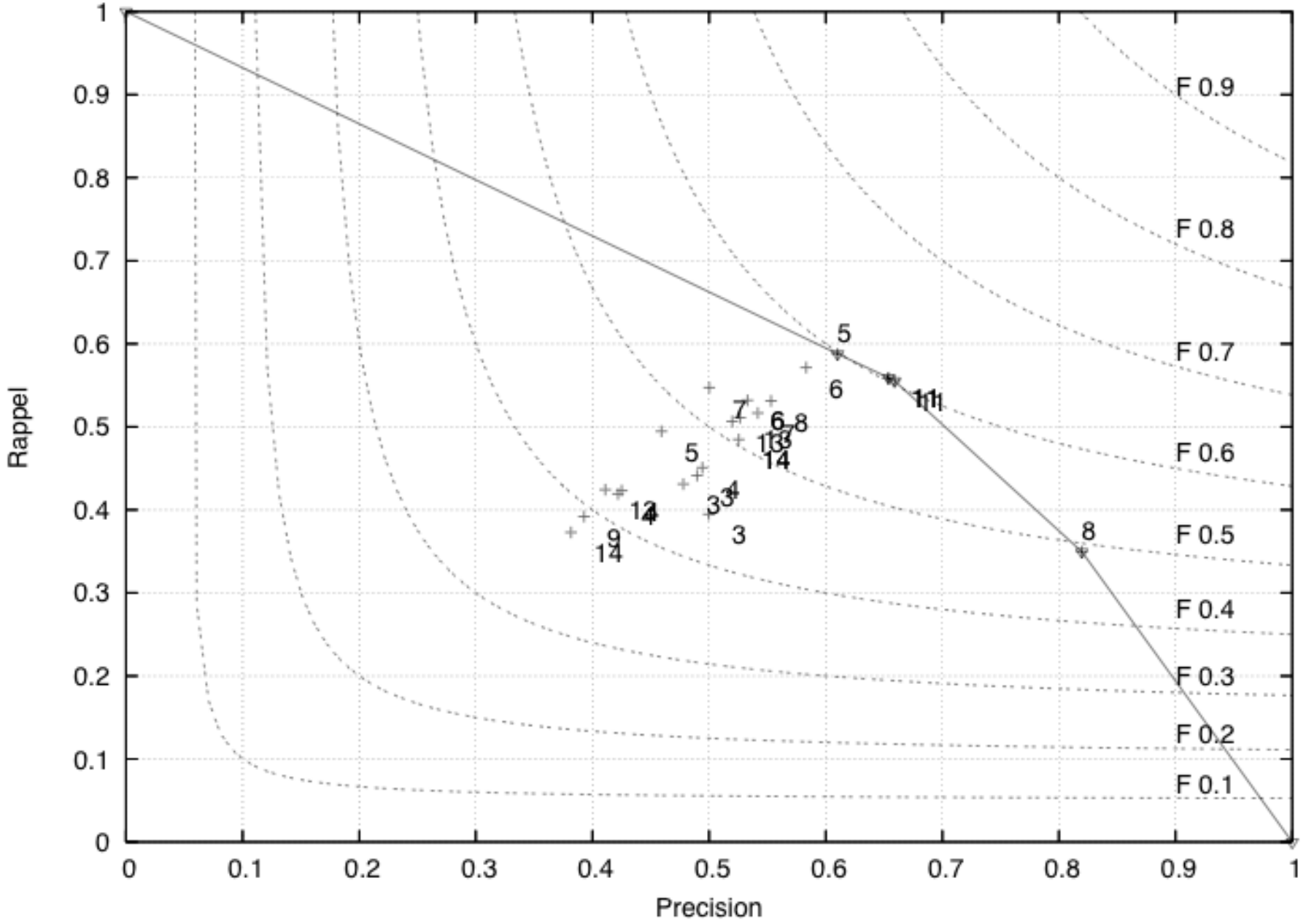
Front de Pareto (jeux vidéos)



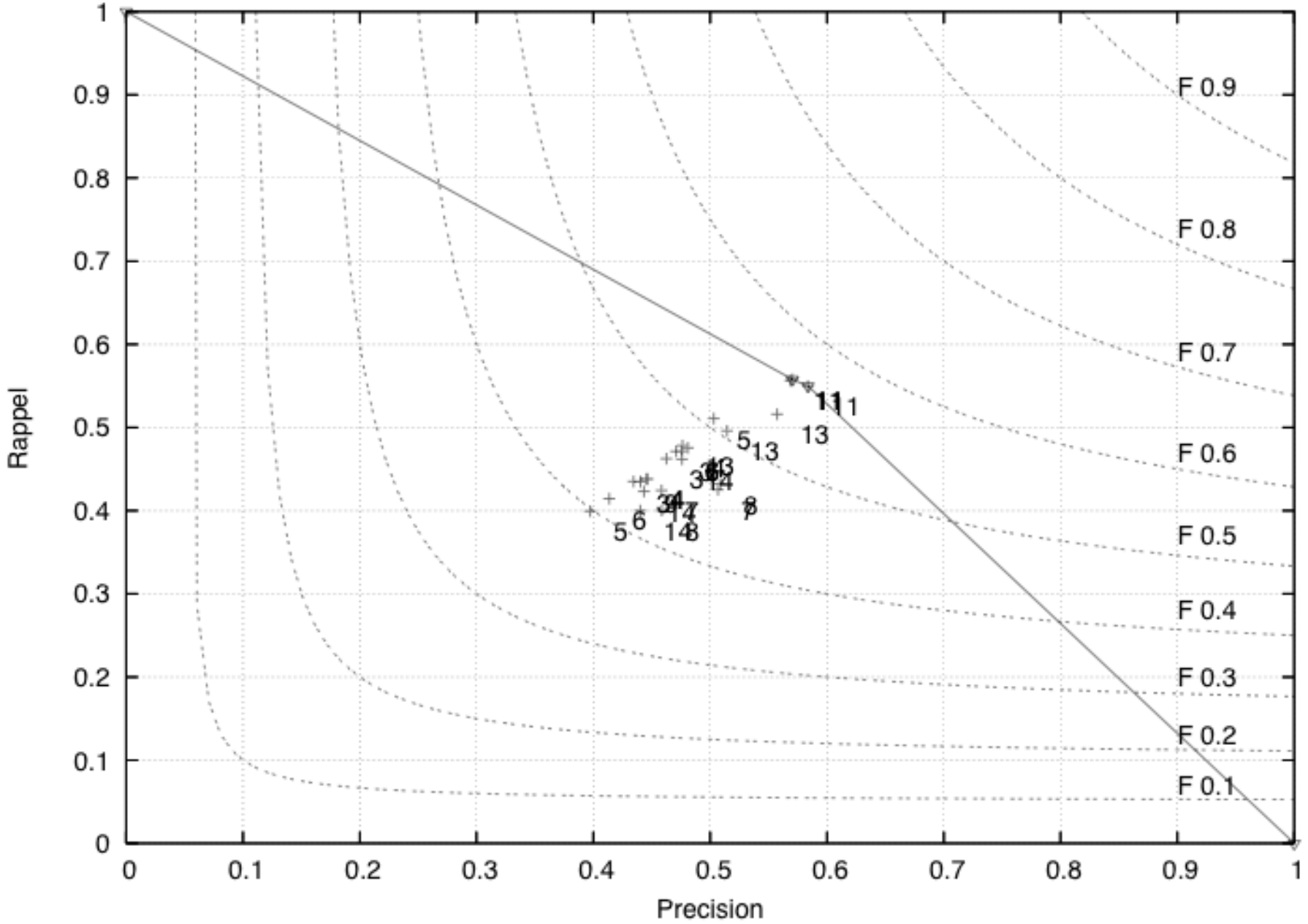
Front de Pareto (débats parlementaires)



Front de Pareto (À voir, à lire)



Front de Pareto (relectures)



Méthodes d'attribution d'une valeur d'opinion

- **Processus constitué de 2 étapes**
  - **Représentation du texte**
  - **Classification**
- **Utilisation de méthodes symboliques …**
  - **Traitements linguistiques**
  - **Vocabulaire d'opinion**
  - **Termes du domaine : focus / attracteur**
- **… et probabilistes / statistiques**
  - **Critères de discrimination des traits**
  - **Classifieurs**

Représentation du texte

- **Extraction de segments pertinents de texte**
  - **Paragraphes (introduction, conclusion)**
  - **Mots liés par une relation d'opinion**
  - **Segment autour d'un mot *attracteur***
- **Lemmes et n-grammes de lemmes**
- **Pondération/sélection des termes d'opinion à partir d'un vocabulaire d'opinion**
- **Critères statistiques ou probabilistes de discrimination**
  - **Tf*idf, gain d'information**

- **Classifieurs *classiques***
  - **SVM, arbres de *décision, régression logistique, réseau de neurones, k plus proches voisins, bayes, similarité …***
- ***Sommation sur les scores attribués aux termes du texte***
- ***Méthodes hybrides***
  - ***Vote entre plusieurs classifieurs***
  - ***Complémentarité***

Merci de votre attention !