Atelier de clôture DEFT2011

Présentation et résultats

Cyril Grouin[1] Dominic Forest[2]
Patrick Paroubek[1] Pierre Zweigenbaum[1]
[1]LIMSI–CNRS [2]EBSI/Université de Montréal

1er juillet 2011 – Montpellier
Atelier TALN

Présentation

Introduction

- Septième édition du défi DEFT ;
- Deux tâches en français :
  - identification de l'année de publication d'un article de presse (continuité DEFT2010) ;
  - appariements résumés/articles scientifiques (nouvelle tâche).

Présentation

Tâche 1 : Diachronie

- Objectif : identification de l'année de publication d'un extrait d'article de presse ;
- Sept journaux français (1800-1944) : *J. des Débats*, *J. de l'Empire*, *J. des Débats politiques et littéraires*, *La Croix*, *La Presse*, *Le Temps*, *Le Figaro*.
- Source : `gallica.bnf.fr`

Présentation

Tâche 2 : Appariements

- Objectif : appariements résumés/articles scientifiques ;
- Six revues francophones : *Etudes internationales, Revue des sciences de l'éducation, Philosophiques, Anthropologie et société, Etudes littéraires, Méta.*
- Source : `www.erudit.org`

Présentation

Calendrier 2011

- **Lancement**
  - 6 janvier et 16 février : appel à participation ;
  - 25 janvier : ouverture des inscriptions ;
- **Déroulement**
  - 21 février : accès aux données d'entraînement ;
  - 4/10 avril : phase de test (fenêtre de trois jours au choix) ;
  - 12 avril : diffusion des résultats aux participants ;
- **Rédaction**
  - 27 avril : remise des articles ;
  - 3 mai : notification ;
  - 9 mai : version finale ;
- **Atelier de clôture** : 1er juillet.

Participants

CHArt (Paris) : YV Hoareau, M Ahat, S Fouchal, C Peterman, D Medernach ;

EBSI (Montréal) : R Boley ;

FBK (Trento) : S Tonelli, E Pianta ;

GREYC (Caen) : G Lejeune, R Brixtel, E Giguet ;

INAOE (Mexico) : F Sánchez-Vega, E Villatoro-Tello, A Juárez-González, L Villaseñor-Pineda, M Montes-y-Gómez, L Meneses-Lerín ;

IRISA (Rennes) : Ch Raymond, V Claveau ;

LIMSI (Orsay) : A García-Fernandez, AL Ligozat, M Dinarelli, D Bernhard ;

LORIA (Nancy)/**LASELDI** (Besançon)/**Diatopie** (Paris) : M Cadot, S Aubin, A Lelu ;

LUTIN (Paris), 2 équipes :

- A El Ghali ;
- D Devatman Hromada.

UCL (Louvain-la-Neuve) : Y Bestgen ;

UPF (Barcelone) : H Saggion.

Tâche 1. Diachronie

Objectif

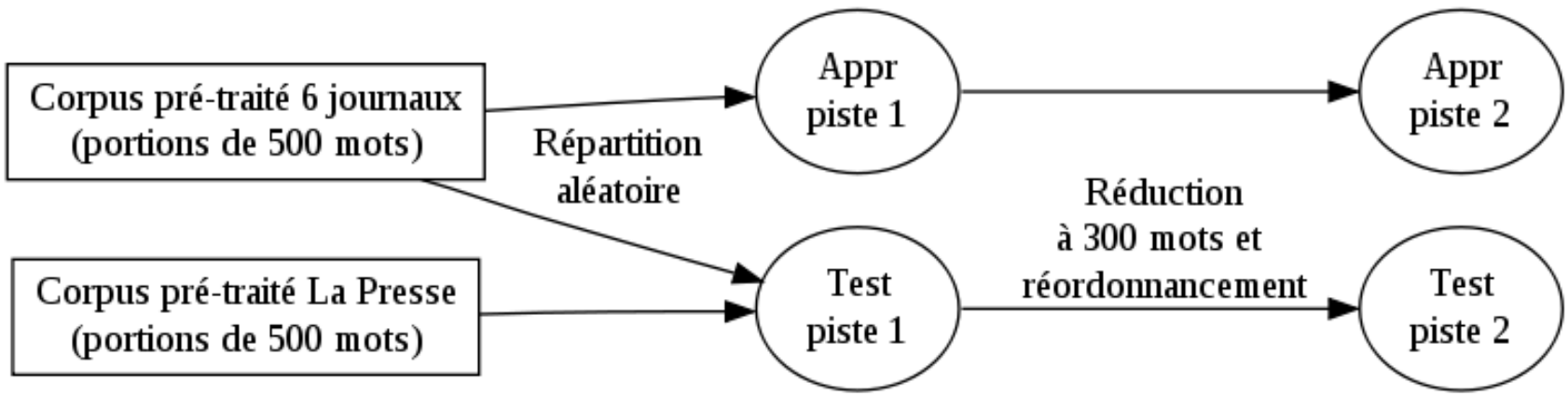
- Identifier l'**année** de publication d'un extrait de journal (DEFT2010 : identification de la **décennie** de publication : problème d'évaluation des documents aux frontières) ;
- Sur une période d'un siècle et demi (1800–1944) ;
- Parmi sept journaux français (+2 par rapport à DEFT2010) ;
- Deux pistes :
  - Portions de 500 mots (nouveau par rapport DEFT2010) ;
  - Portions de 300 mots.
- Hypothèse : *une portion de texte plus importante permet d'améliorer les résultats.*

Constitution des données

- Portail Gallica (BNF) : démarche de numérisation de la presse ancienne (1800–1944) avec reconnaissance des caractères.
- Récupération des journaux numérisés en version texte (*J. des Débats*, *J. de l'Empire*, *J. des Débats politiques et littéraires*, *La Croix*, *La Presse*, *Le Temps*, *Le Figaro*).
- Conservation des deux premières pages uniquement (éviter les pages de programme du théâtre ou de résultats de la bourse).
- Élimination des segments contenant des caractères inutilisés à l'état brut en français : `~ ^ & *` ;
- Segmentation en portions de 500 mots – après suppression des césures –, possiblement à la frontière de deux articles.
- Anonymisation des années (1857, mais ni "!8b2", ni "!92i").

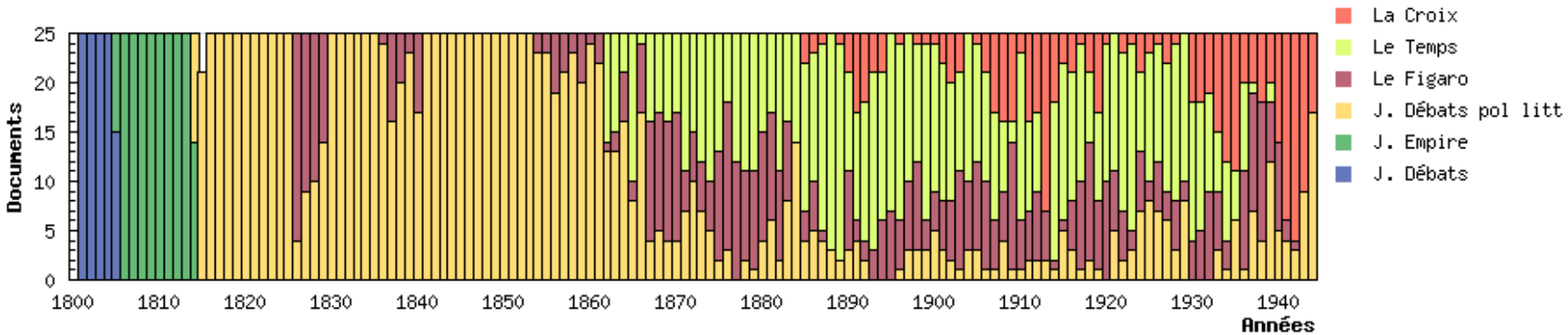
Constitution des données

- Répartition aléatoire des portions de 500 mots en corpus d'apprentissage (60%) et de test (40%) :
  - Seuil : 42 doc/année (apprentissage = 25, test = 17) ;
  - Réservation des articles de *La Presse* pour le corpus de test.
- Pour chaque corpus, réduction des portions à 300 mots et réordonnancement : mêmes documents dans l'apprentissage et le test des deux pistes, dans un ordre différent.

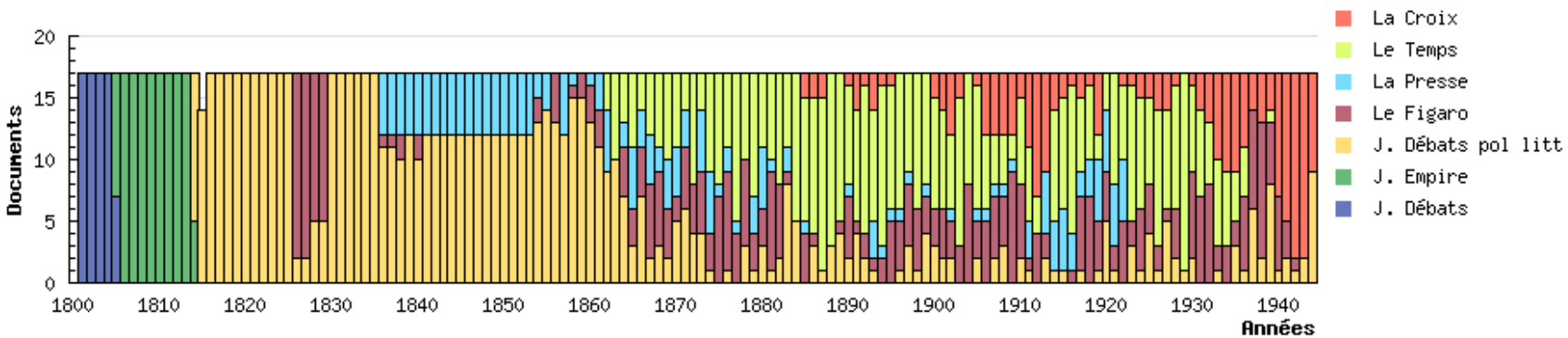


Nombre de documents par année et par journal

- Corpus d'apprentissage



- Corpus de test



Mesures d'évaluation

- DEFT2010 : évaluation binaire (décennie correcte/incorrecte) ;
- DEFT2011 : évaluation sur une fenêtre de 15 ans autour de l'année de référence :
  - Définition d'un gain : similarité entre l'année prédite et l'année de référence ;
  - Plus l'année prédite par le système est proche de la référence, plus le gain sera élevé (fonction gaussienne) ;
  - Score officiel : moyenne des gains sur l'ensemble du corpus ;
  - Si plusieurs hypothèses : gain pondéré par la confiance.

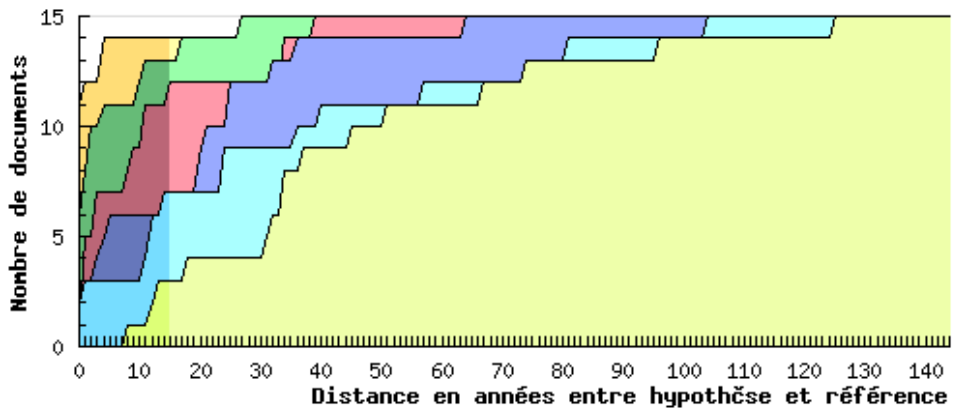
Distance	0	1	2	3	4	5	6	7
Gain	1,000	0,969	0,882	0,754	0,605	0,456	0,323	0,215

Distance	8	9	10	11	12	13	14	15
Gain	0,134	0,078	0,043	0,022	0,011	0,005	0,002	0,001

Tests humains

Procédure

- Test sur les 15 premiers documents du corpus ;
- Forte variation : 0,879 – 0,691 – 0,443 – 0,303 – 0,201 ;
- Score moyen = 0,503 – tirage aléatoire = 0,071 ;
- Méthode : recherche d'entités nommées pour datation ;



Tests automatiques

Procédure

- Adaptation aux données 2011 d'un outil utilisé en 2010 (LIA) : *Isciboost* : algorithme de boosting et arbres de décision ;
- Travail sur les données de test DEFT2011 ;
- Deux types de sortie : décennies (2010) et années (2011) ;

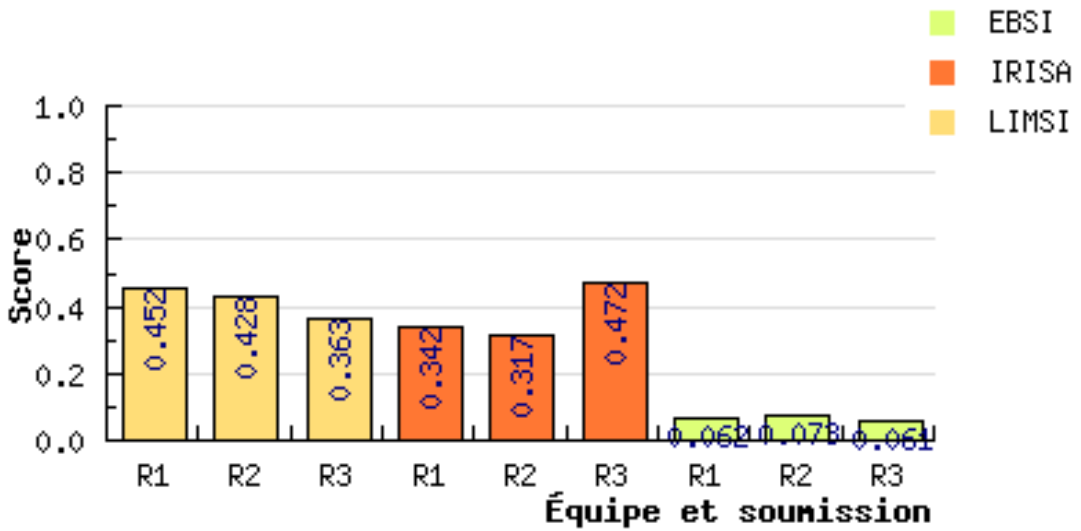
	Décennies		**Années**	
Évaluation	Piste 1	Piste 2	Piste 1	Piste 2
Sans confiance	0,236	0,287	0,140	0,167
Avec confiance	—	—	0,109	0,108

Résultats des participants

	Piste 1			**Piste 2**		
Équipe	Run 1	Run 2	Run 3	Run 1	Run 2	Run 3
EBSI	0,062	0,073	0,061	0,069	—	—
IRISA	0,342	0,317	**0,472**	0,266	0,285	**0,430**
LIMSI	0,452	0,428	0,363	0,378	0,374	0,358
LUTIN-a	(0,098)	(0,108)	(0,100)	0,113	(0,117)	(0,081)
Moyenne	*0,332*			*0,247*		
Médiane	*0,452*			*0,358*		
Écart-type	*0,225*			*0,183*		
Variance	*0,051*			*0,033*		

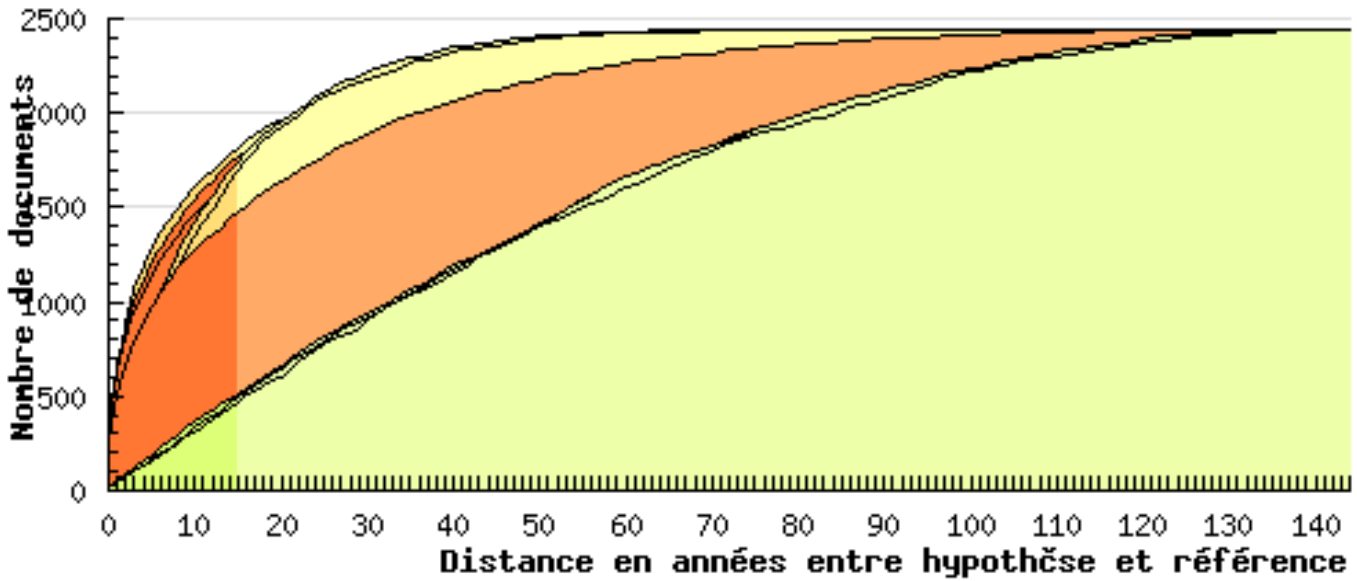
Résultats des participants

Résultats des participants sur la piste 1 :



Résultats des participants

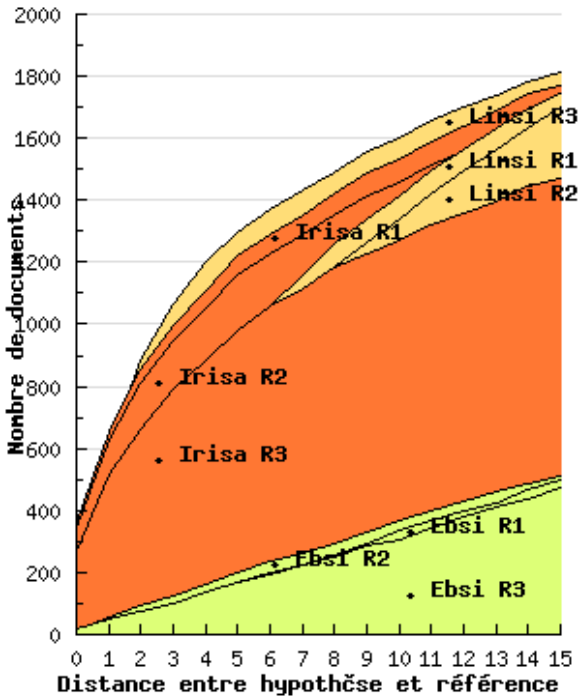
Résultats des participants sur la piste 1 :



- IRISA (orange), LIMSI (jaune), EBSI (vert).

Résultats des participants

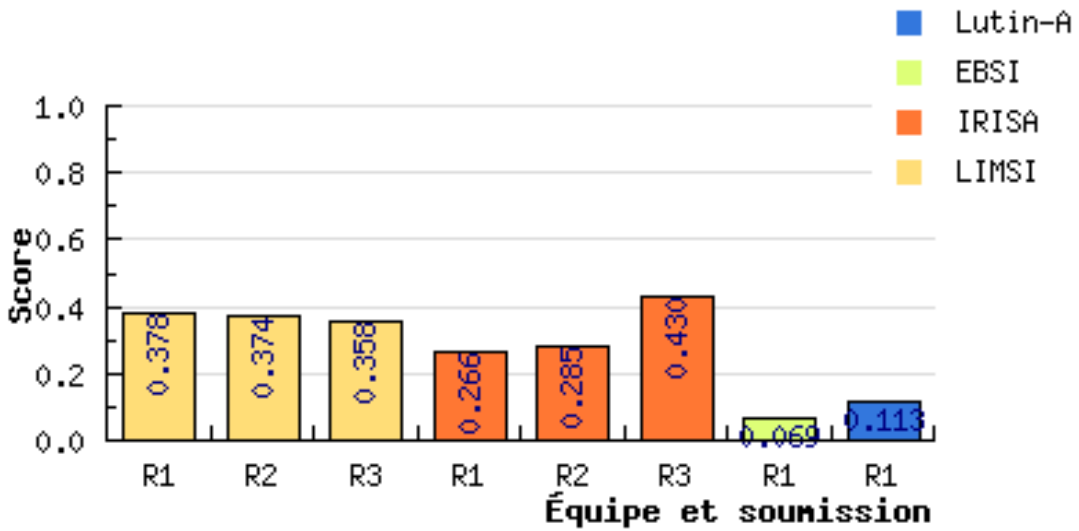
Résultats des participants sur la piste 1 (zoom sur 15 ans) :



- IRISA (orange), LIMSI (jaune), EBSI (vert).

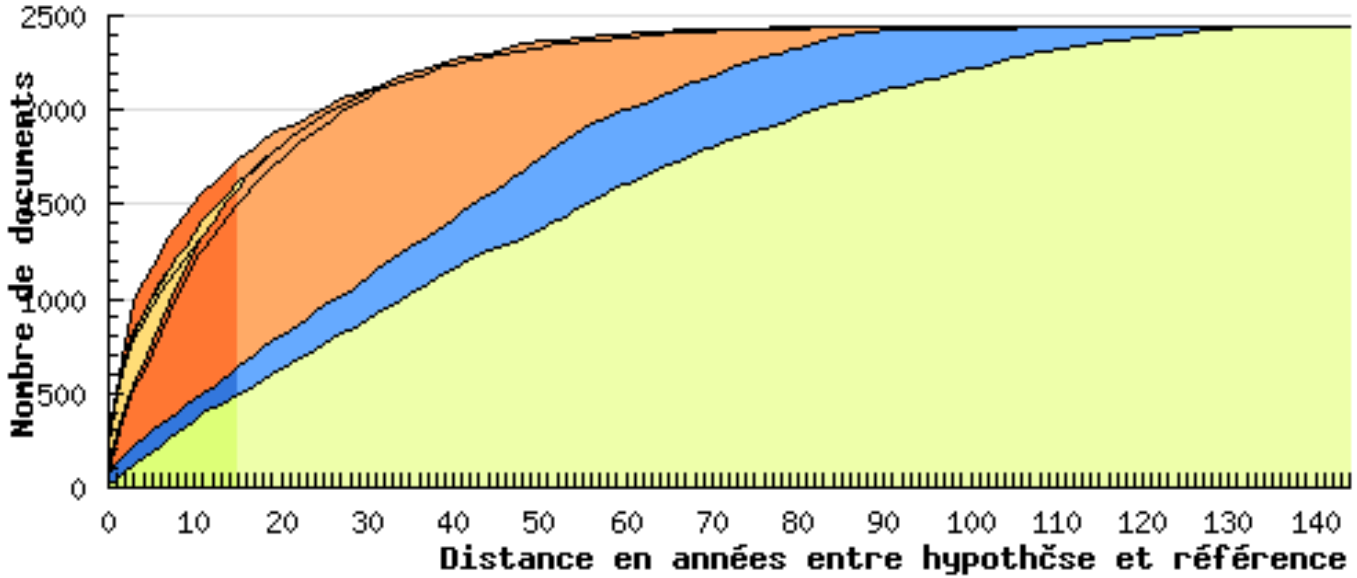
Résultats des participants

Résultats des participants sur la piste 2 :



Résultats des participants

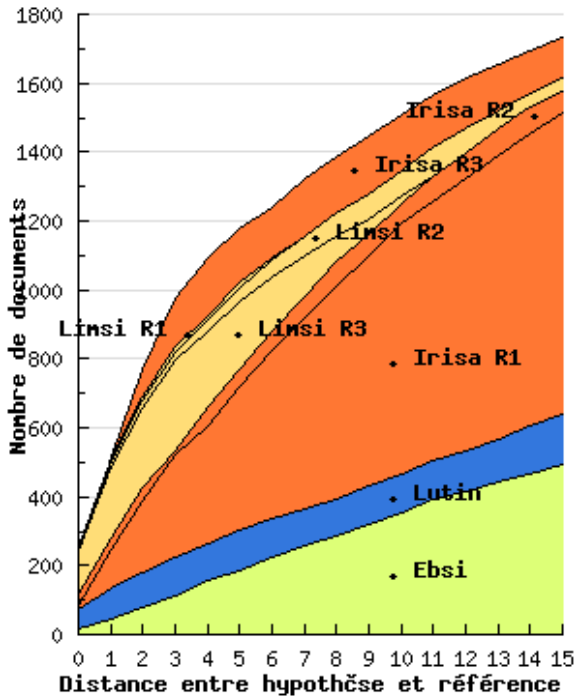
Résultats des participants sur la piste 2 :



- IRISA (orange), LIMSI (jaune), LUTIN-a (bleu), EBSI (vert).

Résultats des participants

Résultats des participants sur la piste 2 (zoom sur 15 ans) :



- IRISA (orange), LIMSI (jaune), LUTIN-a (bleu), EBSI (vert).

Tâche 2. Appariements

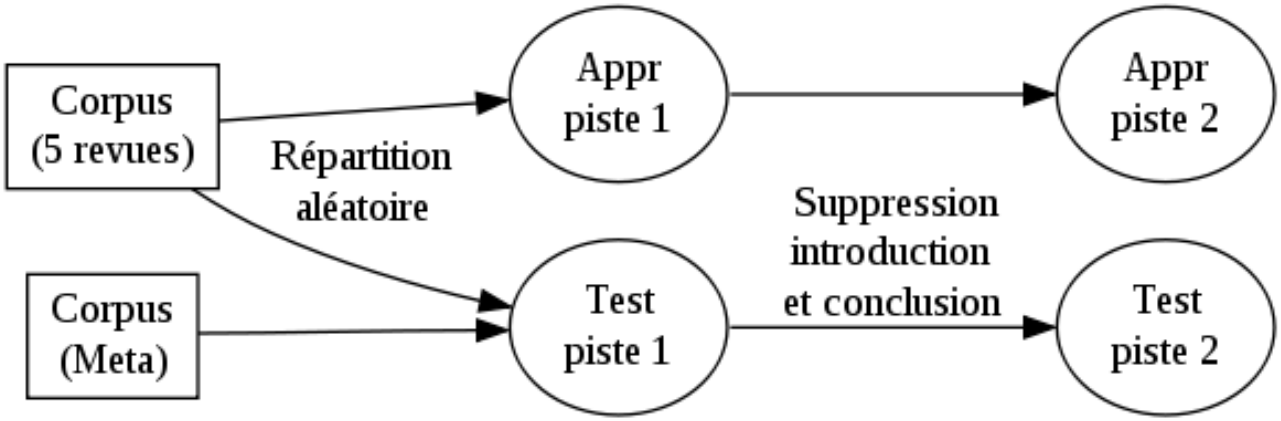
Objectifs

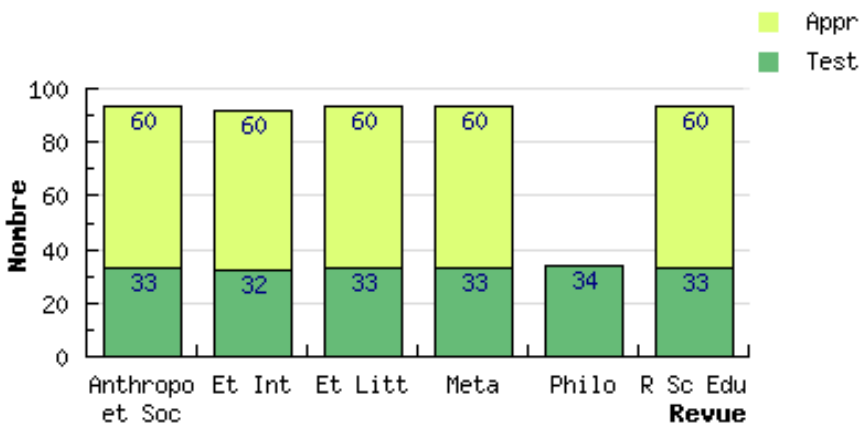
- Apparier un article scientifique avec son résumé ;
- Parmi des revues francophones en SHS (Humanités) ;
- Deux pistes :
  - Piste 1 : appariement résumé/article complet ;
  - Piste 2 : appariement résumé/article sans introduction ni conclusion.
- Hypothèse : *des éléments de l'introduction et de la conclusion sont repris pour générer le résumé.*

Constitution des données

- Plateforme `Erudit.org` : consortium universitaire québécois ;
- Récupération des revues disponibles, sélection de six revues :
  - *Anthropologies et Société*
  - *Études Internationales*
  - *Études littéraires*
  - *Meta*
  - *Philosophiques*
  - *Revue des Sciences de l'Éducation*
- Réservation des articles de *Meta* pour le corpus de test ;
- Apprentissage = 60 articles/revue, test $\simeq$ 30 articles/revue.

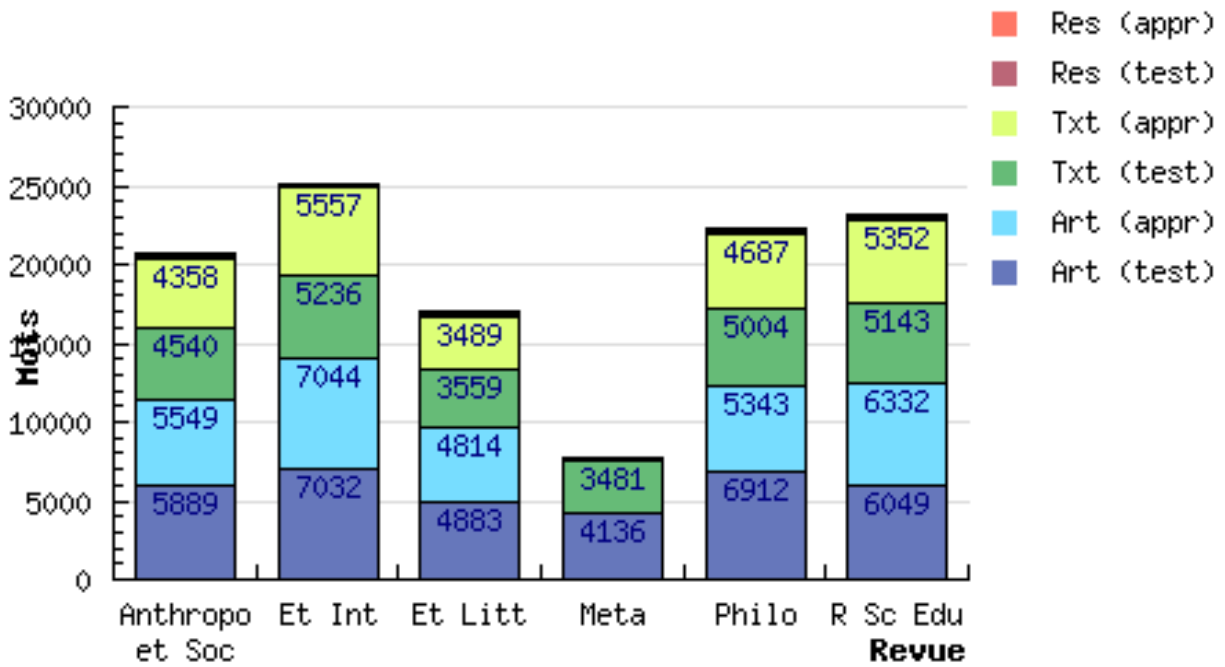
Constitution des données





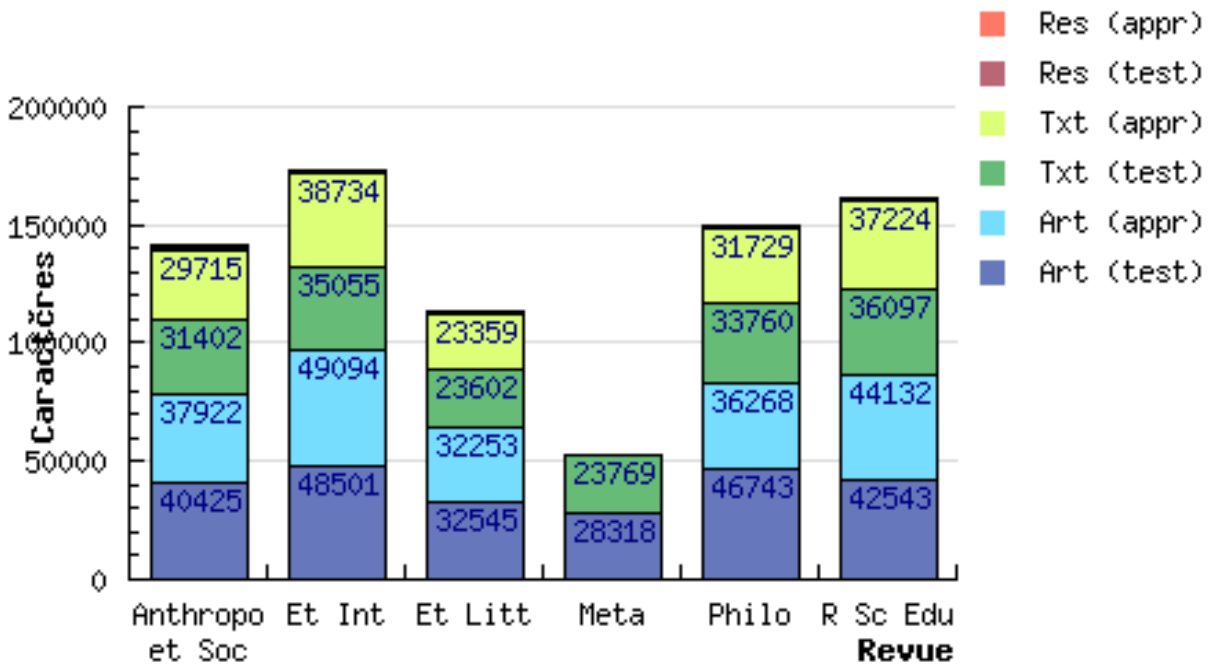
Constitution des données

Nombre moyen de mots par article :



Constitution des données

Nombre moyen de caractères par article :



Mesures d'évaluation

- Évaluation binaire : l'appariements résumé/article est correct (1 point) ou pas (0 point) ;
- Score officiel : moyenne des points obtenus ;
- Si indice de confiance, pondération des points par la confiance indiquée par le système.

Tests humains

Procédure

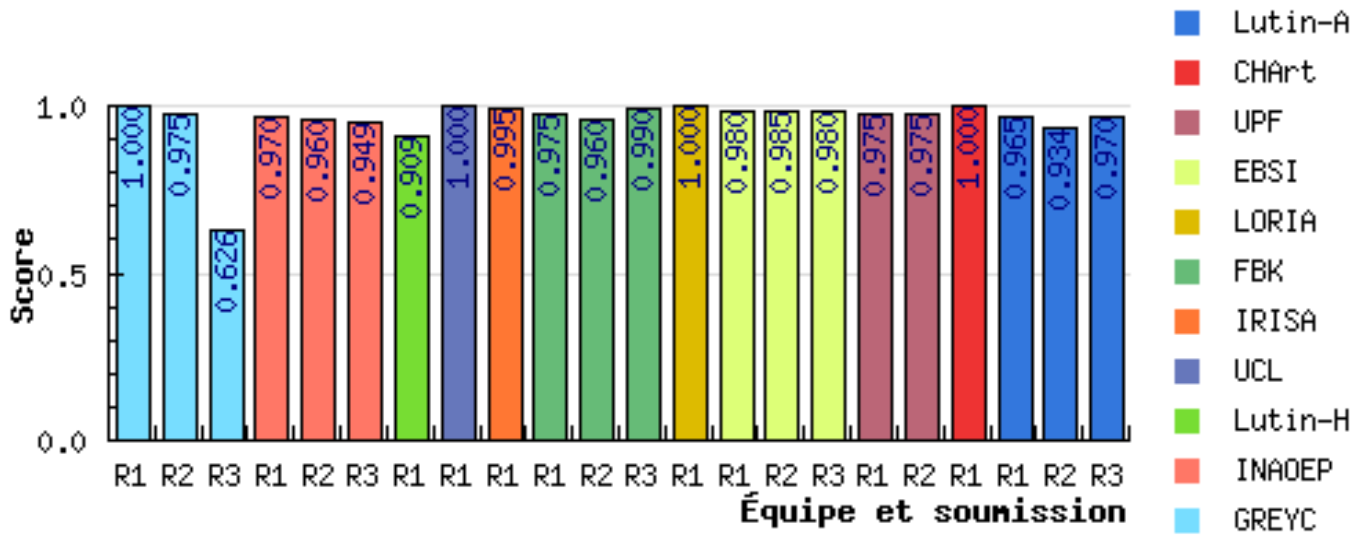
- Test sur 15 couples résumé/articles de deux revues (*Anthropologie et Société*, *Études internationales*) ;
- Score maximum pour chaque évaluateur : tâche (trop ?) facile ;

Résultats des participants

	Piste 1			Piste 2		
Équipe	Run 1	Run 2	Run 3	Run 1	Run 2	Run 3
CHART	**1,000**	—	—	0,995	—	—
EBSI	0,980	0,985	0,980	0,954	0,954	—
FBK	0,975	0,960	0,990	0,964	0,934	0,964
GREYC	**1,000**	0,975	0,626	0,959	0,482	—
INAOE	0,970	0,960	0,949	0,904	0,848	0,858
IRISA	0,995	—	—	0,990	—	—
LORIA	**1,000**	—	—	**1,000**	—	—
LUTIN-a	0,965	0,934	0,970	0,919	0,883	0,873
LUTIN-d	0,909	—	—	0,873	—	—
UCL	**1,000**	—	—	**1,000**	—	—
UPF	0,975	0,975	—	0,959	0,959	—
Moyenne	*0,981*			*0,956*		
Médiane	*0,990*			*0,959*		
Écart-type	*0,027*			*0,042*		
Variance	*0,001*			*0,002*		

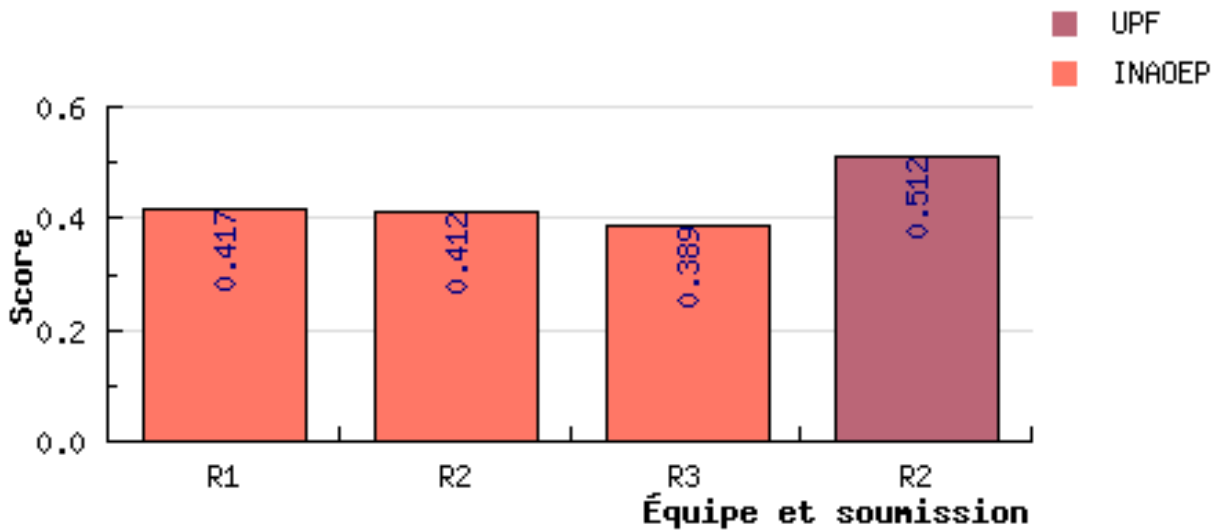
Résultats des participants

Résultats des participants sur la piste 1 :



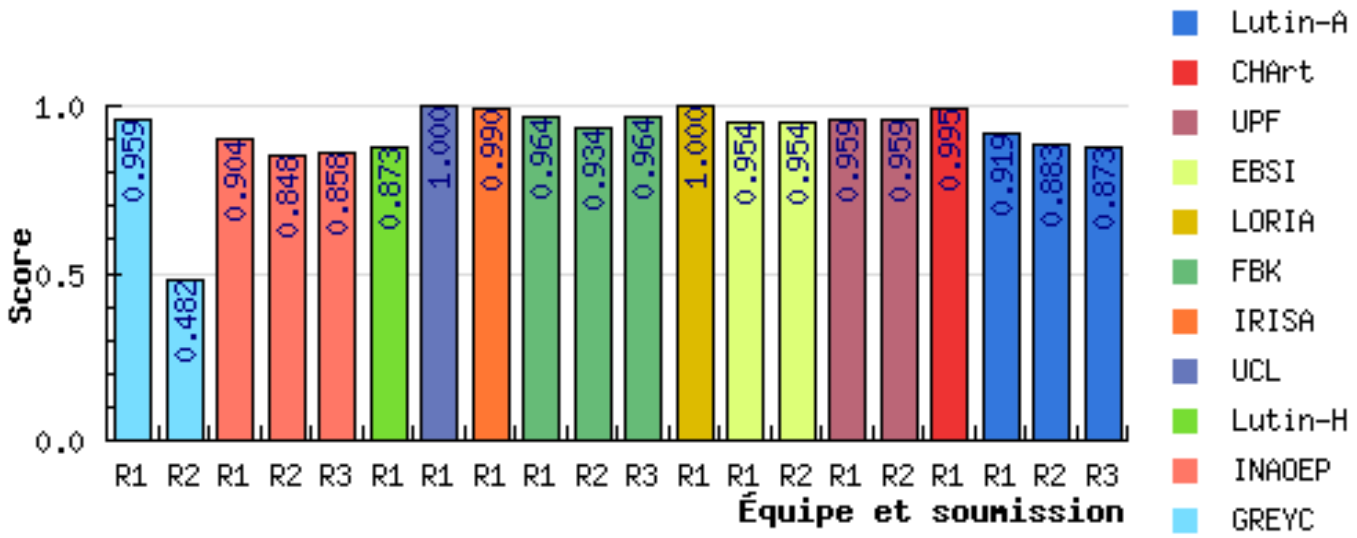
Résultats des participants

Résultats des participants sur la piste 1 (avec prise en compte du score de confiance) :



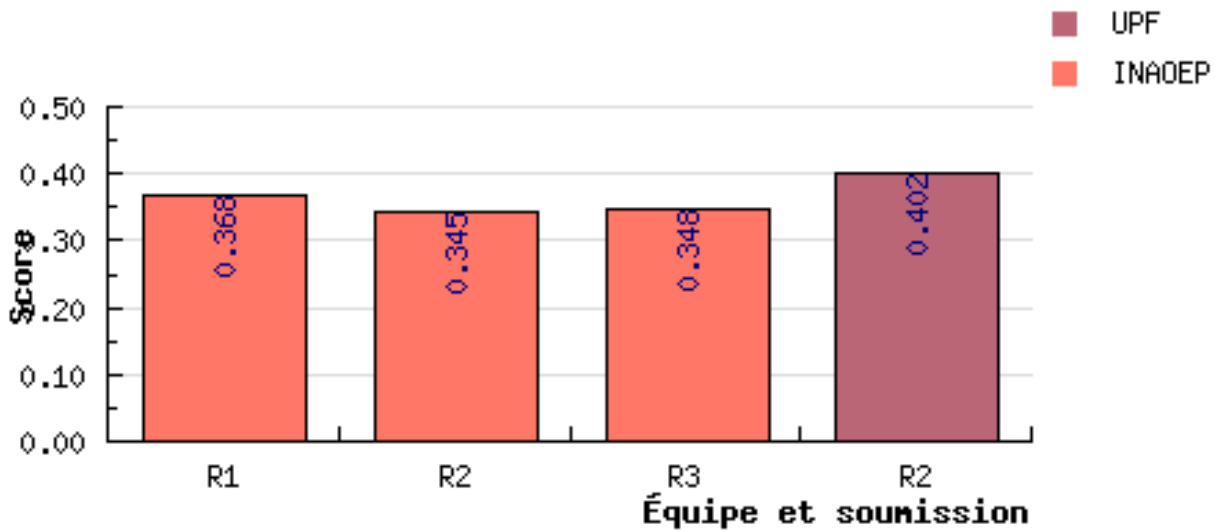
Résultats des participants

Résultats des participants sur la piste 2 :



Résultats des participants

Résultats des participants sur la piste 2 (avec prise en compte du score de confiance) :



Conclusion

- Tâche 1. Diachronie :
  - Tâche difficile mais résultats meilleurs en 2011 qu'en 2010 : F-mesure moyenne = 0,193 (2010) → 0,247 (2011) ;
  - Résultats meilleurs sur les portions de 500 mots (F = 0,332) que sur celles de 300 mots (F = 0,247) ;
  - Nombre élevé de classes et qualité moyenne des documents.
- Tâche 2. Appariements :
  - Tâche trop facile : F-mesure moyenne = 0,981 ;
  - Résultats meilleurs en utilisant les textes complets (0,981) que les textes sans introduction ni conclusion (0,956).

Questions ?