ESTER

Campagne d'évaluation des systèmes d'indexation automatique d'émissions radiophoniques en français.

G. Gravier, J-F. Bonastre, E. Geoffrois,
S. Galliano, K. Mc Tait, K. Choukri

`http://www.afcp-parole.org/ester`

AFCP DGA/CTA ELRA/ELDA

Contexte

- **projet Technolangue EVALDA**
- **financement du Ministère de la Recherche**
- **Organisation**
  - AFCP – responsable des aspects scientifiques de la campagne
  - DGA/CTA – aspects scientifiques, évaluation, fournisseur de données
  - ELDA/ELRA – fournisseur de données, diffusion des ressources

Historique

- **Evaluations NIST/DARPA**
  - Transcription : ATIS, Hub 4, Hub 5, RT
  - Locuteur : SRE, RT
  - Extraction d'information : ACE, TDT
- **En France**
  - évaluation AUPELF des systèmes de dictée vocale (ARC B1)
  - ESTER

Objectifs

- promouvoir une dynamique d’évaluation
- faire un état des lieux
- produire des corpus accessibles pour la recherche
- promouvoir de nouveaux axes de recherche
- inciter les initiatives de collaborations entre différentes branches de la communautés parole et language

Calendrier (janvier 2003 $\longrightarrow$ janvier 2005)

- **Test à blanc (phase 1, terminée)**
  - de juin 2003 à janvier 2004
  - corpus de taille restreinte
  - 10 laboratoires participants
- **Campagne d'évaluation (phase 2, en cours)**
  - de avril 2004 à janvier 2005
  - corpus de taille plus conséquente
  - inscription ouverte jusqu'à la date de diffusion des données de test

Tâches

- **Trois tâches**
  - T — transcription
  - S — segmentation
  - E — extraction d'information
- **Transcription**
  - produire une transcription orthographique
  - catégorie temps réel

Tâches (suite)

- **Segmentation**
  - segmentation en événements sonores — détecter les zones du document de la parole contenant un événement sonore donné
  - segmentation en locuteurs — segmenter le document en tour de parole en regroupant les zones correspondant au même locuteur
  - suivi de locuteur — détecter les zones du document où un locuteur donné est présent
  - suivi interactif de locuteur — idem que suivi de locuteur avec la possibilité de poser des questions à un expert

Tâches (suite)

- **Extraction d'informations (prospectif)**
  - détection d'entités nommées — détecter dans le document sonore les occurences d'entités identifiés (noms, dates, événements, etc)
  - segmentation de document — segmenter le document en section cohérentes par rapport au contenu
  - détection thématique — associer à chaque section du document un ensemble d'indexes thématiques parmi un ensemble de thèmes possibles
  - question / réponse — répondre à une question formulée en langage naturel à partir des documents disponibles

Ressources acoustiques

source	phase 1		phase 2		
	train/dev	test	train/dev	non-trans	test
France Inter	19h40/2h40	2h40	8h/2h	300h	2h
France Info	–	–	8h/2h	1000h	2h
RFI	11h/2h	2h	8h/2h	500h	2h
RTM	–	–	18h/2h	100h	2h
« surprise »	–	–	–	–	2h
total	40h		50h	2000h	10h
période	1998–2000		2003	2004	2004

Ressources textuelles

- **Le Monde**
  - article du journal Le Monde
  - 1987 $\longrightarrow$ 2003
  - 300M mots
- **MLCC**
  - transcriptions des débats du parlement européen
  - 5,5M mots
- **transcription du corpus de développement**

Ressources dérivées

Ressources produites par les participants susceptible d'être diffusées :

- phonétisation du corpus de développement
- alignement phonétique
- transcription automatique
  - mesure de confiance
  - N-meilleurs hypothèses
  - graphes de mots
- modèles (acoustiques, langage)

Conclusion

LES ORGANISATEURS SONT OUVERTS A TOUTES PROPOSITIONS CONCERNANT DES TRAVAUX SUR LES DONNEES ESTER