

Évaluation des Systèmes de Transcription Enrichie d' Émissions Radiodiffusées

Guillaume Gravier

MINISTÈRE DE LA DÉFENSE



<http://www.afcp-parole.org/ester>



Plan de l'exposé

1. Un aperçu de la campagne ESTER
 - tâches
 - existant
 - objectifs
2. L'organisation
 - organisateurs
 - élaborations des protocoles
 - participants
 - résultats
3. Création de corpus et valorisation
4. Bilan

Tâches évaluées

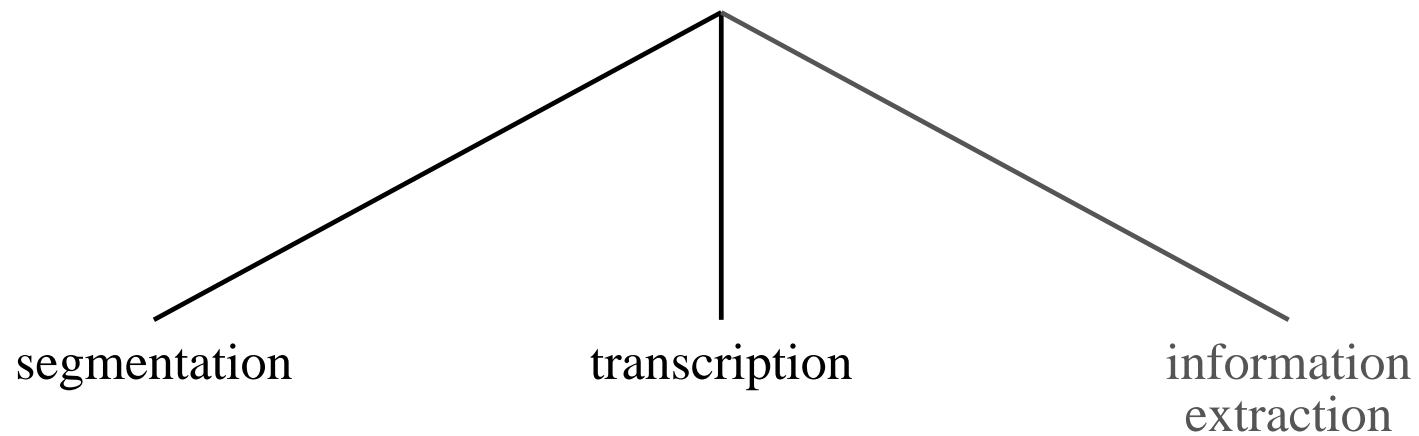
Broadcast News Indexing

(French Language)

Tâches évaluées

Broadcast News Indexing

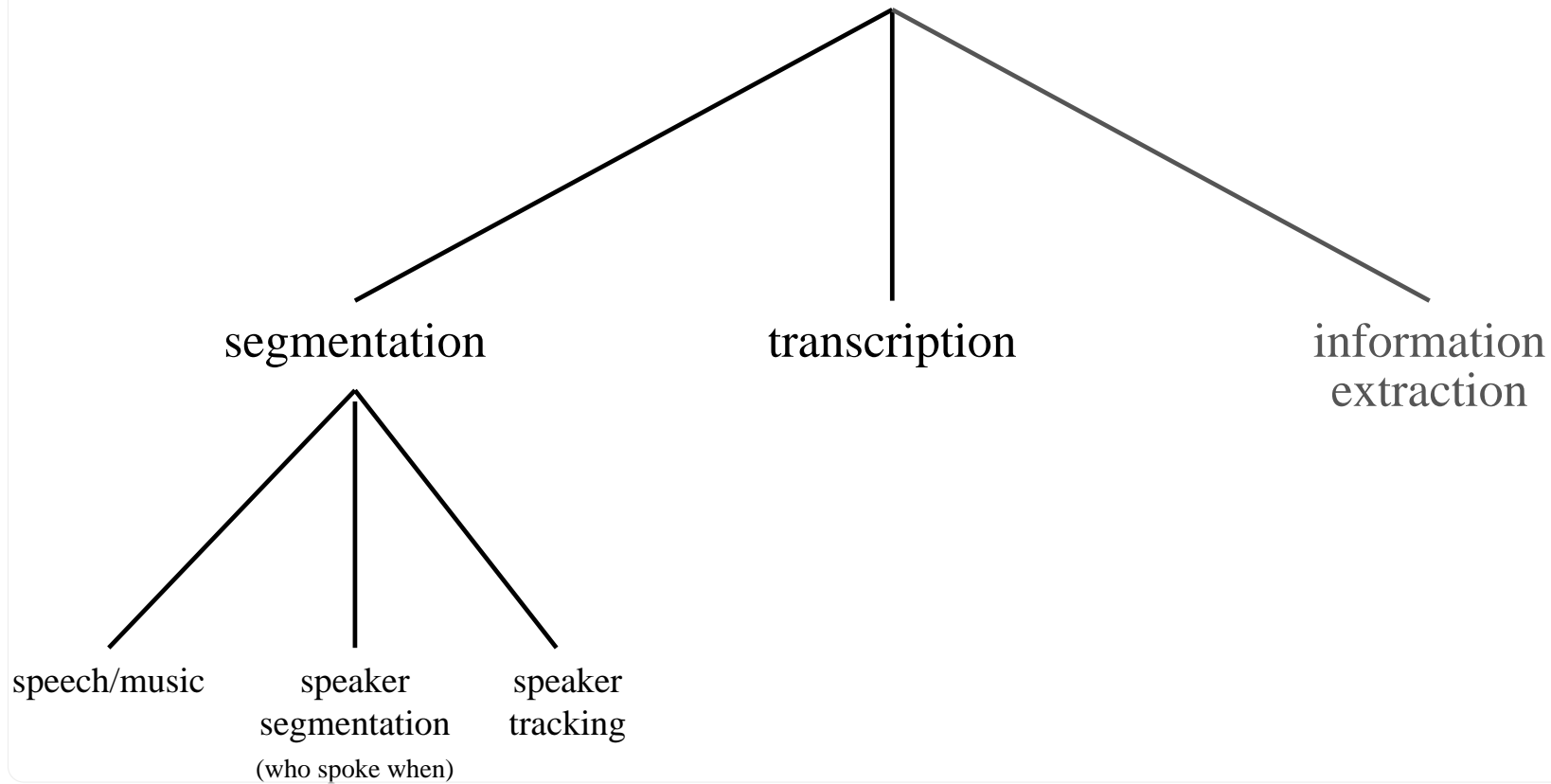
(French Language)



Tâches évaluées

Broadcast News Indexing

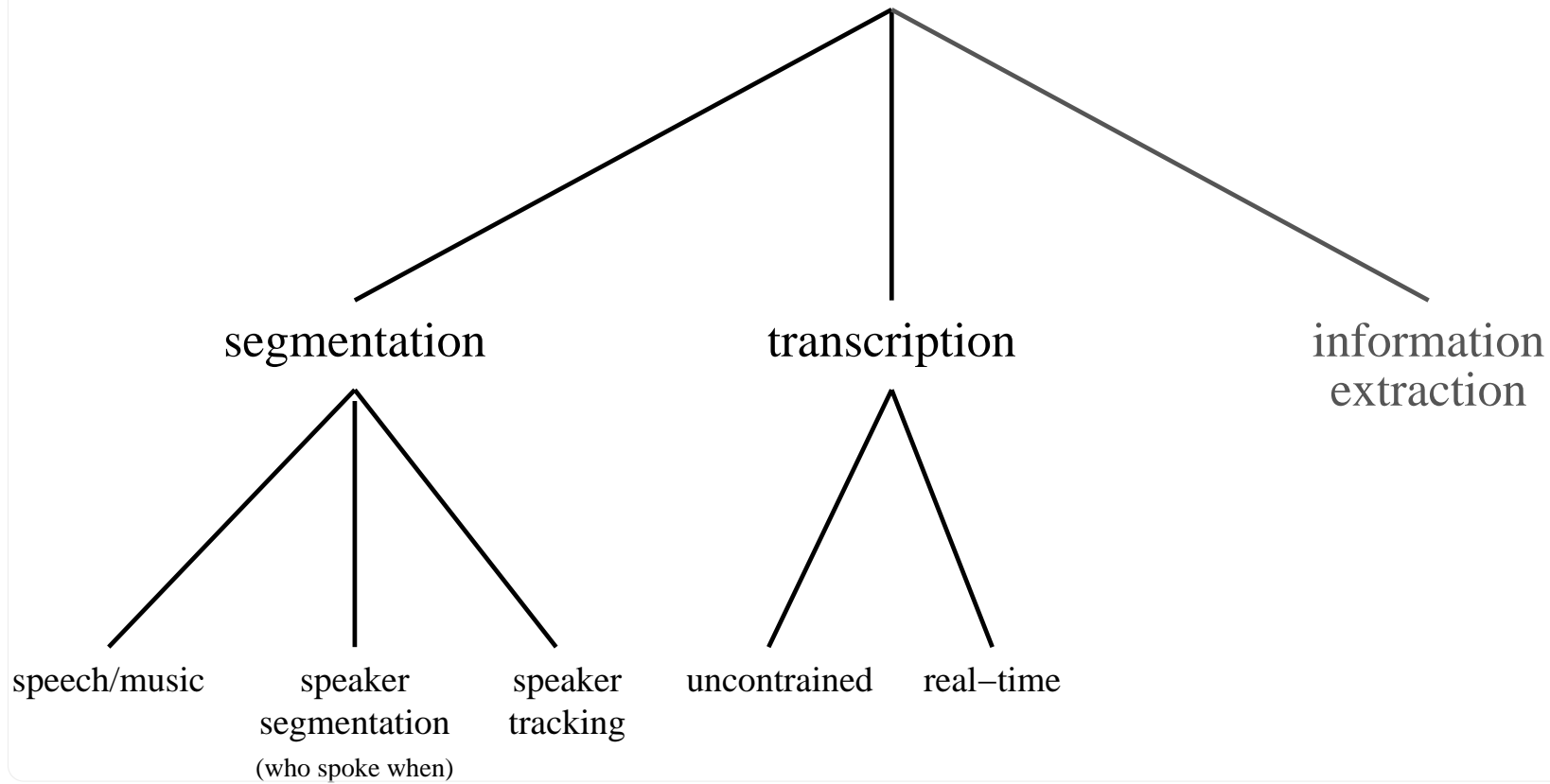
(French Language)



Tâches évaluées

Broadcast News Indexing

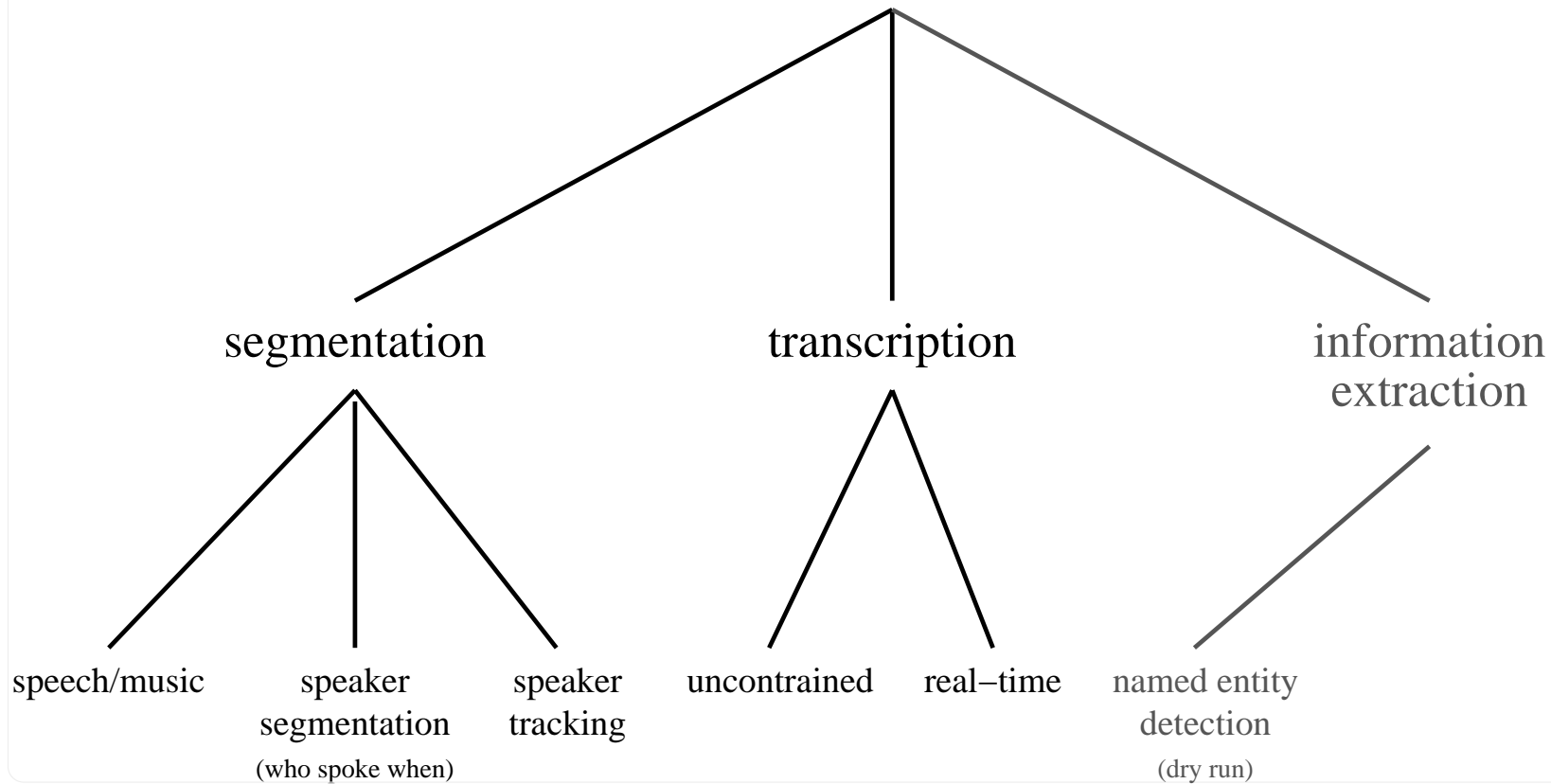
(French Language)



Tâches évaluées

Broadcast News Indexing

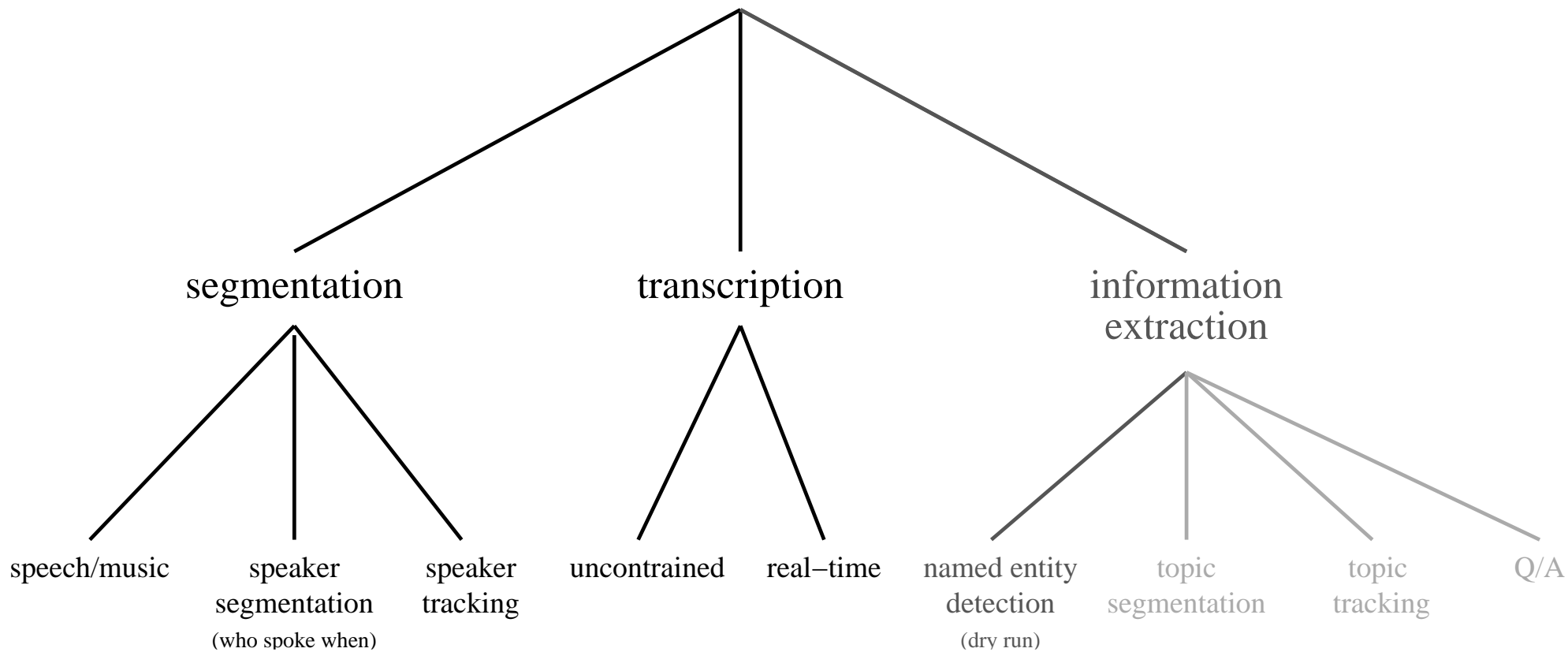
(French Language)



Tâches évaluées

Broadcast News Indexing

(French Language)



Existant

Transcription

- 1996 – 1999, DARPA Hub 4 : informations radio et télé
- 2002 – 2006, NIST Rich Transcription : informations radio et télé, conversations téléphoniques, réunions
- 1995 – 1996, AUPELF ARC B1 : parole lue

Existant (suite)

Segmentation

- 1998, NIST Speaker Recognition and Tracking : suivi de locuteurs cibles dans des conversations téléphoniques
- 1996 – 2006, NIST Speaker Recognition : vérification d'identité
- 2002 – 2006, NIST Rich Transcription : détection des tours de parole

Existant (suite)

Extraction d'information

- 1998 – 2004, [Topic Detection and Tracking](#) :
détection, segmentation et classification thématique,
suivi de sujets (sur transcriptions automatiques)
- 1999 – 2002, [NIST Automatic Content Extraction](#) :
détection et suivi des entités nommées
- 2002 – 2006, [NIST TREC Spoken Document Retrieval](#) : recherche de documents
- 1997 – 2000, [EC Research Project THISL](#) :
Thematic indexing in Spoken Language

ESTER par rapport à l'existant

ESTER mélange des tâches bien connues (conventions, métriques, etc.) à des aspects plus exploratoires :

⇒ évaluation de la segmentation parole/musique

⇒ suivi de locuteurs cibles

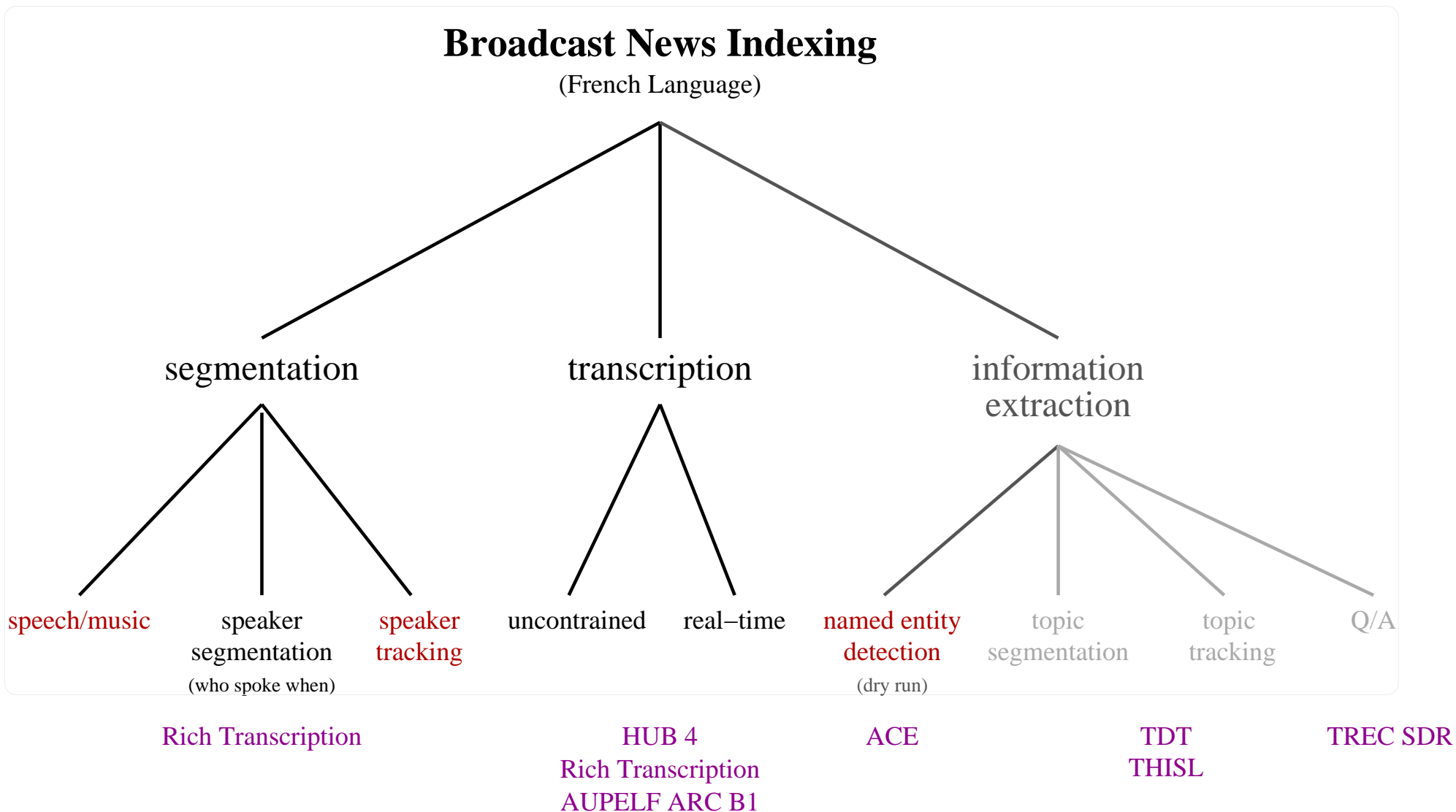
et développe le traitement automatique de la parole en français :

⇒ pas d'évaluation en langue française depuis l'AUPELF !!!!!

⇒ pas de corpus disponibles en français équivalent aux corpus américains

⇒ pas de conventions communément admises pour les entités nommées

ESTER par rapport à l'existant



Objectifs

Les objectifs principaux de ESTER sont donc

- assurer l'accès aux ressources annotées en français pour l'ensemble de la communauté scientifique
- assurer une dynamique de la communauté, notamment par l'organisation de campagnes d'évaluation pérennes
- évaluer les performances des différents systèmes
- et les faire progresser

Les organisateurs

Association Francophone de la Communication Parlée

- animation scientifique
- définition du plan d'évaluation
- valorisation scientifique

Centre d'Expertise Parisien de la DGA

- définition du plan d'évaluation
- production de ressources
- validation des résultats
- animation scientifique

ELRA

- production de ressources
- diffusion des ressources

Déroulement de la campagne (phase 1)

Phase 1 : janv. 2003 à mars 2004

- utilisation de 40h d'émissions radiophoniques annotées par la DGA, disponibles au départ du projet (et déjà à la disposition de quelques participants)
- test à blanc en janvier 2004 (8 participants)
- atelier de présentation des résultats en mars 2004
- enregistrement et annotation de 50h de données supplémentaires

Déroulement de la campagne (phase 2)

Phase 2 : mars 2004 à mars 2005

- protocoles et métriques d'évaluation revus à la lumière de l'expérience acquise lors du test à blanc
- utilisation des 40h + 50h pour le développement
- campagne d'évaluation en janvier 2005 (13 participants)
- atelier de présentation des résultats en mars 2005

Elaboration des protocoles (phase 1)

Tâches fixées lors de la rédaction de la proposition initiale. Protocole et métrique établis en collaboration avec les participants!

- **proposition des protocoles** par les organisateurs du projet
- **discussion avec les participants** potentiellement intéressés
- **finalisation des protocoles** pour le test à blanc
- développement des logiciels de mesure de performances lorsque nécessaire (organisateur)

Elaboration des protocoles (phase 2)

- **identification des problèmes** de protocole et de métrique de la phase 1 avec les participants lors de l'atelier
- **nouvelle proposition** des organisateurs
- **finalisation** des protocoles finaux (processus itératif)
- adaptation des logiciels de mesure de performances
- définition du contenu du corpus de test
- **phase d'adjudication**

Les participants

Participants non financés mais les données restent à leur disposition à l'issu de la campagne (à des fins de recherche).

Différents profils de participants :

- transcription : laboratoires avec une **longue expérience** sur la transcription enrichie (LIMSI, Vecsys Research) vs. laboratoires avec une **expérience en transcription** mais pas sur ce type de tâche (LIA, LORIA, IRISA/ENST, LIUM), voire **pas d'expérience** (CLIPS, IRIT).

Les participants (suite des profils)

- segmentation parole/musique : de **nombreux laboratoires** (avec plus ou moins d'expérience) dont certains n'ayant pas la masse critique pour se lancer dans les autres tâches (Univ. Balamand, LISIF)
- locuteurs : la plupart des laboratoires ont une **expérience des évaluations NIST** sur le locuteur (IRISA, ENST, LIMSI, LIA, CLIPS, FT R&D)
- entités nommées : participants avec une **expérience sur le texte mais pas sur la parole**, voire pas d'expérience

peu de participants hors France !!!!!

Les résultats

laboratoire	TRS	TTR	SES		SRL	SVL	EN*	
			par.	mus.			ref	asr
CLIPS	40,7				27,2			
ENST/TSI	45,4					47,0		
FT R&D			99,1					
IRISA-ENST/INF	35,4		98,9	33,7		84,3		
IRIT	61,9	70,4	98,8	52,7				
LIA	26,7	36,3	99,2	54,8	19,2	66,0	34,1	57,4
LIMSI	11,9				11,5			
LIPN							37,0	
LIUM	23,6		97,4		16,9		39,7	61,2
LORIA	27,6	37,4	97,5					
Univ. Balamand			95,1	26,2				
Vecsys Research		16,9						

Gros progrès entre phase 1 et phase 2 grâce à l'expérience acquise et aux données supplémentaires

(*) résultats préliminaires (non officiels) du test à blanc.

Création de corpus

Diffusion aux participants de ressources existantes

- 40h enregistrées par le CEP (DGA)
- corpus de textes (Le Monde 1987 – 2003, MLCC)

Création de nouvelles ressources

- 60h (50h + 10h de test) de données annotées
- 2 000h de données non annotées

Propriété des corpus créés aux organisateurs.

Création d'outils

Développement de ressources pour la mesure des performances

- logiciels pour les tâches de suivi d'événements (parole/musique, locuteur)
- dictionnaire d'équivalence pour la transcription

Travaux sur les entités nommées (post évaluation)

- définition de conventions strictes d'annotation
- annotation cohérente de l'ensemble du corpus

Propriété des outils créés aux organisateurs.

Valorisation et diffusion

- **publications** présentant la campagne, les résultats, les corpus : LREC 2004, JEP 2004, Interspeech 2005, LREC 2006.
- **diffusion des corpus** et du lot d'évaluation à faible coût pour la recherche (300/2000 euros)
- **promotion** auprès de la communauté Science Humaines et Sociales : organisation d'une journée de rencontre entre organisateurs ESTER, participants à la campagne et laboratoires intéressés par l'utilisation du corpus (principalement linguistes et phonéticiens)
- mise en place d'un **espace public d'échange** permettant d'enrichir le corpus par des ressources dérivées automatiquement : transcriptions, alignements phonétiques, graphes de mots, etc.

Les bons points...

- plusieurs systèmes opérationnels pour les évaluations
- progrès notables pour de nombreux laboratoires
- un corpus public de référence pour les tâches évaluées (et d'autres éventuellement)
- de nombreuses publications des participants : 9 publications liées à ESTER au dernier Interspeech!
- une nouvelle dynamique de la communauté TAP et des liens naissant avec la communauté SHS

Les bons points... (suite)

- contact avec NIST pour certaines tâches
- discussion pour inclure une tâche 'speaker tracking' dans la prochaine campagne NIST Rich Transcription
- contact avec le groupe Broadcast News du COST 278 pour un échange de ressources

... et les moins bons!

- temps de développement (ingénierie) très long \implies un investissement très lourd pour pouvoir faire de la recherche !!!!!
- beaucoup de temps passé à reproduire des choses existantes
- risque d'optimiser les systèmes par rapport à la métrique d'évaluation (recherche de la performance) au détriment d'aspects plus fondamentaux

Il est nécessaire de poursuivre l'effort pour profiter des développements et des progrès déjà effectués...

... sans pour autant passer son temps à évaluer!