

The ESTER Evaluation campaign for the Rich Transcription of French Broadcast News

Guillaume Gravier

Edouard Geoffrois

Kevin Mc Tait

Jean-François Bonastre

Sylvain Galliano

Khalid Choukri

AFCP

DGA/CTA

ELDA

LREC 2004, Lisbon, May 27th

Overview

- Context
- Goals, tasks
- Schedule, resources
- Participants, derived resources
- Perspectives

Context

- NIST evals on English BN (96–) and RT (2001–)
→ *established know-how, protocols & scoring tools*
- Previous French eval on speech dictation (95–99)
→ *expectation of a follow-up*
- BN project at DGA/CTA (98–)
→ *French BN data available*
- Technolangue program (launched in 2002)
→ *budget, opportunity*

Technolanguge

Data & tools

- ...
- ...
- ...
- ...
- ...
- ...
- ...
- ...
- ...
- ...

Evaluation
campaigns

- ESTER
- ...
- ...
- ...
- ...
- ...
- ...
- ...
- ...

Norms &
standards

- ...
- ...

Technology
portal

- Technolanguge.net

Technolanguge

Data & tools

- ...
- ...
- ...
- ...
- ...
- ...
- ...
- ...
- ...
- ...

Evaluation campaigns

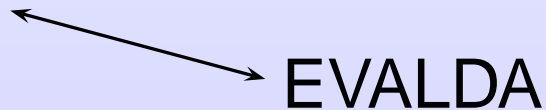
- **ESTER**
- MEDIA
- EvaSy
- EASy
- EQueR
- CESART
- ARCADE II
- CESTA
- ...
- ...

Norms & standards

- ...
- ...

Technology portal

- Technolanguge.net



The ESTER project

Evaluation of spoken language technology for French BN:

- single material: 100h of French (radio) BN
- several tasks: transcription, speaker recognition, IR&E
- NIST-style evaluation campaign

Goals of the project:

- promote evaluation for French
- develop resources, which live up beyond the project
- encourage developments in a collaborative fashion, especially between spoken and written language processing communities
- evaluate system performances

Organizers

- AFCP (French Speech Communication Association)
→ protocols, web site, workshop organization
- DGA/CTA (French defense lab for info processing)
→ protocols, evaluation, data provider
- ELDA (Evaluation and Language Resource Agency)
→ data provider, dissemination

Tasks

abbrev.	description
T / TRS	orthographic transcription
T / TTR	real time transcription
S / SES	sound event tracking
S / SRL	speaker diarization
S / SVL	speaker tracking
S / SIL	interactive speaker tracking
E / EN	named entity detection
E / SD	document segmentation
E / ST	topic detection and tracking
E / QR	information retrieval (question answering)

Schedule

Dry run (January 2004):

- using 40h of acoustic data, subset of the tasks
- 6 sites submitted (5 for transcription)

Official tests (January 2005):

- using 100h of acoustic data, all tasks
- 14 sites registered, registration remains open until test data release

Participants submitting results can keep the data for free (for research purposes) after the project.

Acoustic data

source	train1	train2	non-trans	dev	test
France Inter	25h	8h	~200h	2h	2h
France Info	–	8h	~800h	2h	2h
RFI	15h	8h	~900h	2h	2h
RTM	–	18h	~100h	2h	2h
“surprise”	–	–	–	–	2h
total	40h	42h	~2000h	8h	10h
period	98–00	2003	2004	2004	2004

Text data

- Le Monde
 - Newspaper articles
 - 1987 → 2003
 - 300M words
- MLCC
 - transcriptions of the European parliament debates
 - 5.5M words
- detailed transcriptions of the acoustic data
 - ~1M words

Submissions to dry run

Institution	Tasks		
CLIPS	TRS		SRL
ENST1	TRS		
France Telecom			SRL
IRISA+ENST2	TRS	SES	SVL
LIA	TRS		
LORIA	TRS	SES	

Registrations for official eval

Institution	Tasks					
CLIPS	TRS	TTR	SES	SRL		
DDL			SES	SRL	SVL	SIL
ECL			SES	SRL	SVL	SIL
ENST	TRS		SES			
France Telecom				SRL	SVL	
IRISA	TRS	TTR	SES	SRL	SVL	
IRIT	TRS		SES	SRL		
LIA	TRS	TTR	SES	SRL	SVL	SIL
LIMSI	TRS		SES	SRL		
LIUM	TRS		SES			
LORIA	TRS		SES			
SIS	TRS		SES			
Univ. Balamand	TRS		SES			
Vecsys Research	TRS	TTR	SES	SRL		

Derived resources

Some resources produced by the participants are expected to be disseminated:

- phonetization of the lexicon
- phonetic alignments
- automatic transcriptions
 - confidence measures
 - N-best hypotheses
 - word graphs
- acoustic and language models

Perspectives

- Official tests
- Data release (low cost for research)
- New evaluation campaigns
- Strengthen European and international cooperation

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