Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE

Ministère de la Culture et de la Communication

Ministère de la Recherche

Ministère de l’Economie, des Finances et de l’Industrie

TechnoLangue: «Language Technologies» Action

Joseph Mariani

Director

« Information & Communication Technologies » Department

French Ministry of Research

CSLF Report

- Report produced by CTIL (Chair : A. Danzin) for CSLF (Vice - Chair : B. Cerquiglini) and submitted to Prime Minister in November 2000
- Interministerial meeting on June 26, 2001
- 3 actions:
  - 1. Technological survey and French language processing tools evaluation (Ministry of Research)
  - 2. Promotion of language processing technologies in administration (MCC and MFPRE)
  - 3. Training of professionals in Document Engineering (MEN)

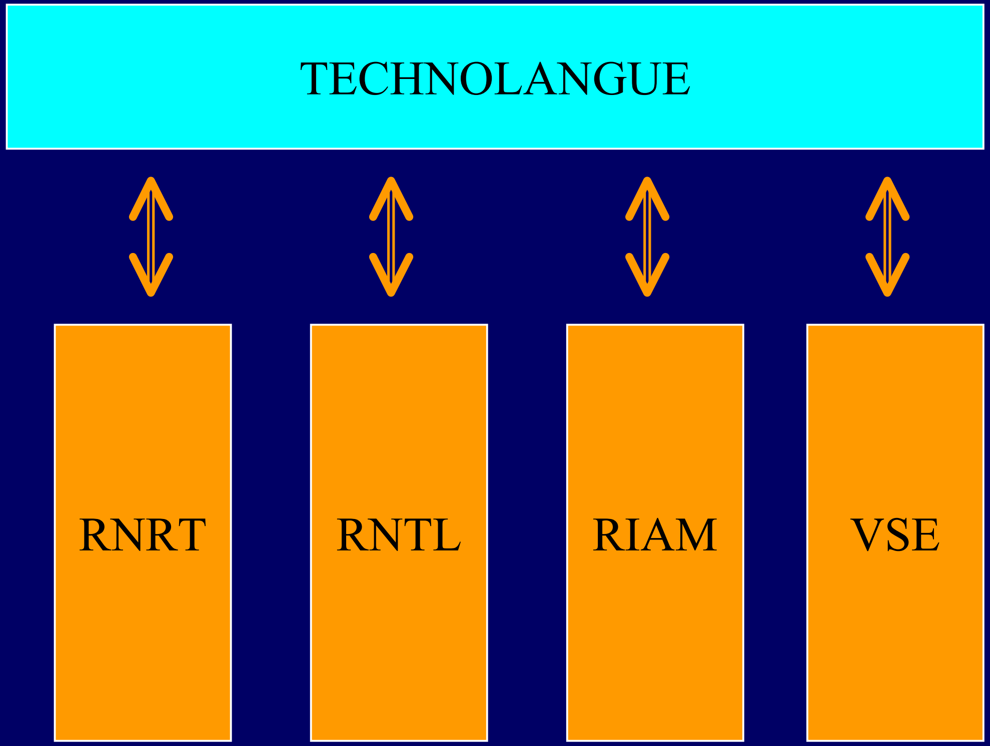
« TechnoLangue » action

- Action 1 (TechnoLangue) launched by MR, MinEFI and MCC
  - Budget <= 4 M€ (2002)
  - First interministerial large program specifically dedicted to (spoken and writtten) language technologies
- Linked to Research and Innovation Technological Networks (RRIT)
- Comparable action in other domains
  - Image Technologies in 2003 ?

RRIT

- Technological Research and Innovation Networks (RRIT)
  - – 4 ICT RRIT: RNRT (Telecommunications), RNTL (Software), RIAM (Audiovisual & Multimedia), (RMNT (Micro & Nano-Technologies)))
  - – Cooperative R& D projects (public research - industry)
  - – 100 M€ total 2002 budget (MR+MinEFI+MCC)
  - – MR : Funded by Technological Research Fund (FRT)
    - 150 M€ : total budget in 2002 (16 RRIT)
    - 38 M€ for 4 ICT RRIT
    - 3.8 M€ for transversal actions, such as TechnoLangue
  - – Also VSE: Technological survey over the Internet (MR)

« TechnoLangue » structure


TECHNOLANGUE

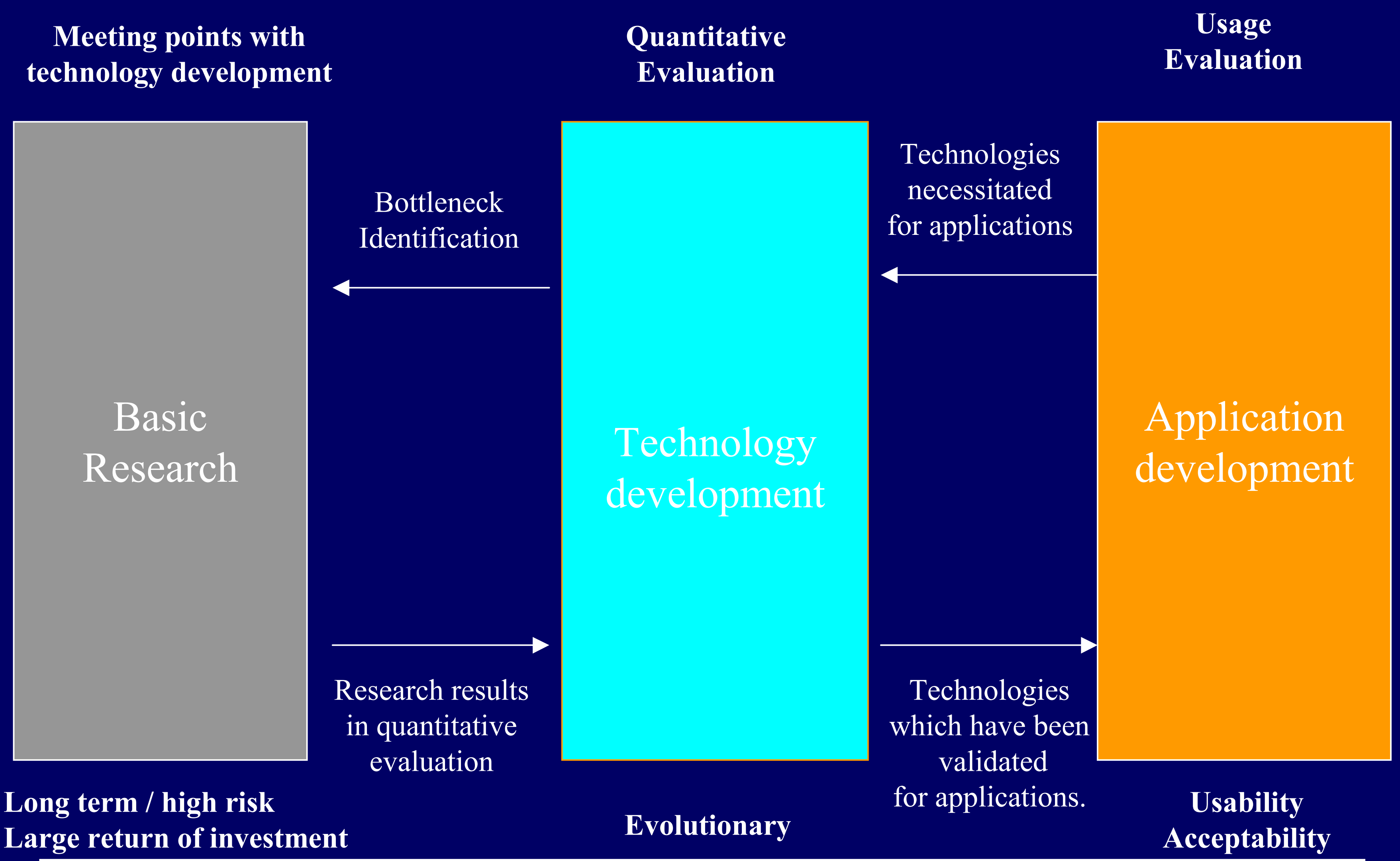
RNRT

RNTL

RIAM

VSE

Infrastructure program to support technological innovation, while existing R&D projects stay with RRIT & VSE



Meeting points with technology development

Quantitative Evaluation

Usage Evaluation

Basic Research

Bottleneck Identification

Technology development

Technologies necessitated for applications

Application development

Research results in quantitative evaluation

Technologies which have been validated for applications.

Long term / high risk
Large return of investment

Evolutionary

Usability
Acceptability

« TechnoLangue » action

- Organization
  - Executive Committee (EC) chaired by C. Fluhr (CEA)
  - Comprising 15 members:
    - 3 RRIT representatives: B. Bachimont (INA - RIAM), C. Sedogbo (Thalès - RNTL), C. Waast (IBM - RNRT)
    - 3 Public research: C. Fluhr (CEA), E. Geoffrois (DGA) P. Paroubek (Limsi-CNRS)
    - 5 Industrials: K. Choukri (ELDA), B. Normier (Lingway), J.-J. Rigoni (Elan Informatique), F. Segond (Xerox) + C. Sorin (FT R&D)
    - 4 Administrations: S. Chaudiron (MR), J. Mariani (MR), D. Malbert (MCC), J. Mathieu (MinEFI)
  - Good balance between research & industry - written/spoken

« TechnoLangue » action

- Install a User Committee
  - Ministry of Foreign Affairs
    - Automatic translation, multilingualism…
  - Ministry of public administration
    - Simplification of the administrative language...
  - Ministry of National Education
    - Training technologies, language traning...
  - …

« TechnoLangue » Call

- 4 meetings of the Executive Committee
- A Call for Proposals with 4 parts
  - Part 1: Language resources
  - Part 2: Evaluation
  - Part 3: Norms & standards
  - Part 4: Technological survey
- Submitted to the ICT RRIT for comments
- Calendar:
  - Launched April 15, 2002
  - Deadline : May 31 / June 10 (Electronic) - June 17 (Paper)
  - Results : July 19, 2002

« TechnoLangue » Call

- International cooperation
  - Similar national programs
    - EU Countries (Italy, Norway, Germany, Greece, The Netherlands, Switzerland…)
    - Prepare the construction of the European Research Area
    - USA, Japan, South Africa…
  - Cooperation mechanisms
    - foreign entities may participate in the project
    - proposals should be written in French
    - financing from own funds

Part 1: Language Resources

- Stimulate the production and the distribution of language resources for :
  - answering minimal needs (*Basic LAnguage Resource Kit*) for the french language ;
  - promoting resources reusabilty ;
  - supporting research ;
  - helping industrial applications development ;
  - decreasing the cost of entering the sector for new comers
- Should include the French language, eventually in connection with other languages

Part 1: Language Resources

- Spoken and written data :
  - oral corpus, pronunciation lexicons, etc.
  - databases for speech synthesis ;
  - monolingual and multilingual text corpus (parallel, comparable...) ;
  - lexicons, terminology, grammars,...
  - Lexical semantic resources : ontologies, thesauri,...
  - Multimodal corpus,...etc
- Basic sofware tools :
  - morphosyntactic taggers, syntactic parsers, semantic tools,
  - teminology extractors,
  - language identifiers,
  - corpus annotations tools,
  - lemmatizers,… etc.

Part 1: Language Resources

- Encourage and facilitate the use of those resources
  - Putting them in new (young) user hands
  - Same approach as for GUIs : “VUIs”
  - Language Technology Kits with “User’s guide”
    - Distribution towards specialized education entities (NLP, Document Engineering…) and more largely towards training centers (Universities, Technical Universities, Engineering schools...)
    - While insuring a feedback from experience
  - Open Source software economical model

Part 2: Evaluation

- 3 areas :
  - Technology evaluation
  - Application evaluation
  - Evaluation methodologies

Part 2: Evaluation

- Technology evaluation
  - Organization of comparative evaluation campaigns for technologies presently not covered by european or international programs, or with a complementary approach
  - Includes the production of the data necessary for the evaluation, in a monolingual, multilingual or crosslingual context
  - Scientific and industrial interest of the evaluation should appear (large enough number of participants)
  - The projects must define the evaluation methodology and justify the practical organization aspects

Part 2: Evaluation

- Application evaluation
  - The objective is to develop evaluation mehodologies for industrial or pre-industrial products
  - The methodologies may result in “toolboxes”, also regrouping user-oriented methodologies and protocols, or in test software packages
  - The methodologies should be generic (class of applications)
  - The proposals should demonstrate the project economical and industrial interest, and the modalities of the distribution of the “toolboxes”

Part 2: Evaluation

- Evaluation methodologies
  - Improve the present evaluation methodologies
  - Identify new (quantitative and qualitative) approaches for already evaluated technologies :
    - socio-technical and psycho-cognitive aspects
    - cognitive modeling of evaluation
  - Identify protocols for new technologies and applications
    - Virtual Reality, Multimodal interaction, Language on the Internet...

Part 3: Standards

- Support the participation of French actors in normalization and standardization bodies
  - Presently weak participation of French actors in normalization and standardization bodies
  - Of strategic importance
  - Variety of places where the normalization activities are taking place : official or non-official committees, forums, projects,...

Part 3: Standards

- Actions:
  - Support the creation of consortia to reinforce the french presence in various bodies (ISO, CEN, W3C,...)
  - Help the share of efforts among French participants
  - Identify a topic and ensure a permanent participation in all related bodies : character sets, exchange format, phonetic alphabet transcription, etc.
  - Necessity of articulating the project with French bodies already implied : AFNOR, W3C French Chapter,...

Part 4: Survey

- Part 4 - Install an information survey
  - Create a portal on Language Engineering in order to give access to :
    - panorama of the industrial and technological offer
    - state-of-the-art in science and technology
    - identification of language resources
    - identification of technological bottlenecks
    - a list of Call for Proposals
    - a presentation of the market key numbers
    - an information on norms and standards (with Internet links)
  - Should be linked with existing sites (Euromap,...)

Conclusions

- Launch a large national program on Language Technology (TechnoLangue)
- In the perspective of installing a permanent infrastructure for LR, Evaluation, Standards and Survey
- Hope that it can participate in the construction of the European Research Area
- And articulates well with international activities