



---

*Ministère de la Culture et de la Communication*  
*Ministère de la Jeunesse, de l'Éducation Nationale et de la*  
*Recherche*

*Ministère de l'Économie, des Finances et de l'Industrie*

**Action TechnoLangue**

**«Technologies de la langue »**

**J. Mariani - Directeur DT/A3**

**S. Chaudiron - Chargé de Mission DT/A3**

---

# Rapport du CSLF

---

- Rapport réalisé par le CTIL (Président : A. Danzin) pour le CSLF (Vice - Président : B. Cerquiglioni) et remis au Premier Ministre en Novembre 2000
- Réunion interministérielle à Matignon le 26 juin 2001
- 3 actions:
  - Veille technologique et évaluation des outils de traitement de la langue française (MR)
  - Développement des usages du traitement informatique de la langue française (MCC et MFPRE)
  - Formation de professionnels en ingénierie documentaire (MEN)

# Action « TechnoLangue »

---

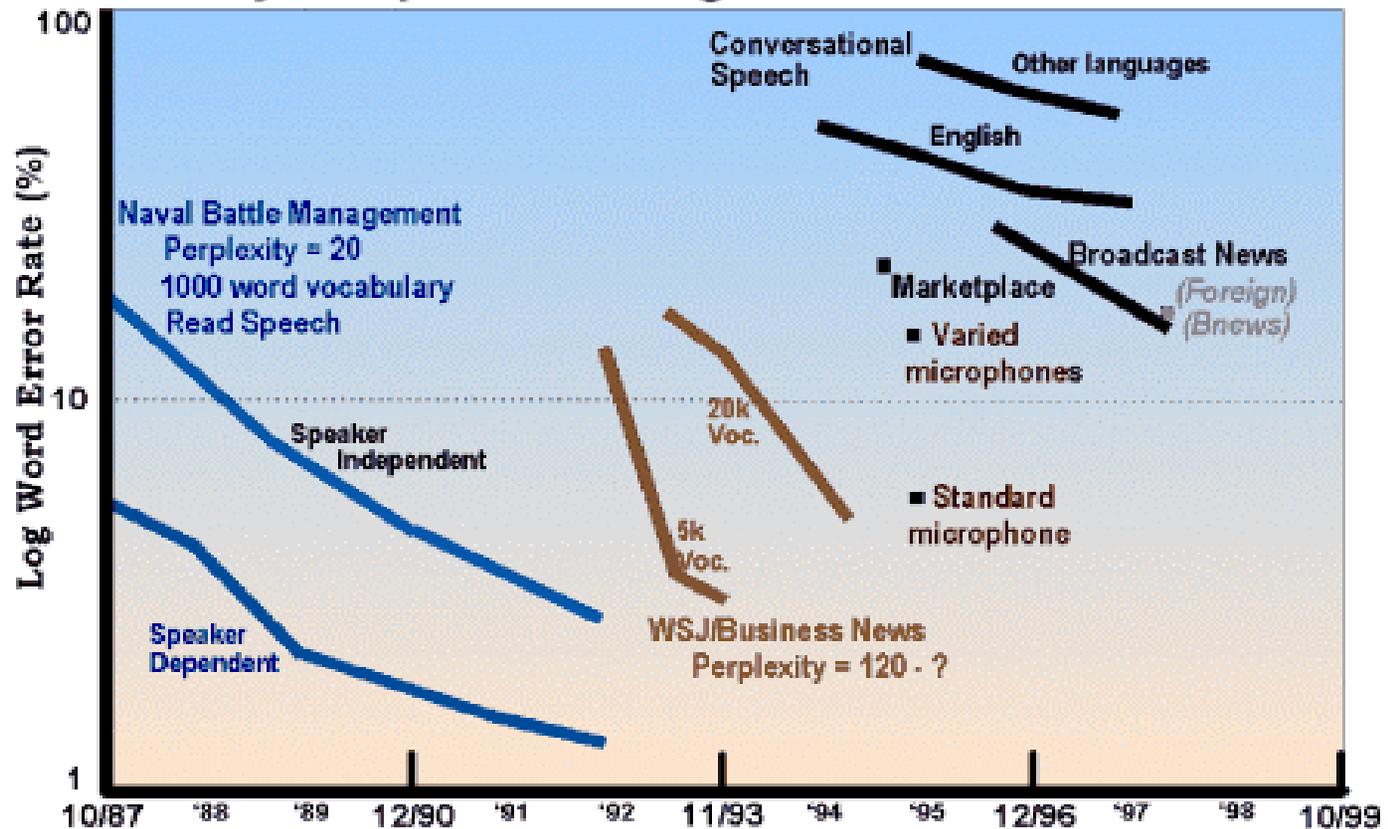
- Lancée à la demande du Premier Ministre par le MR, le MinEFI et le MCC
  - Doté d'un budget  $\leq 4$  M€ (année 2002) ;
  - Premier programme interministériel de grande ampleur spécifiquement dédié aux technologies de la langue (écrite et orale)
  - Programme coopératif (associe laboratoires publics et industriels, au moins 3 partenaires (MR))
- Articulée avec les Réseaux de Recherche et d'Innovation Technologiques (RRIT) concernés

# Contexte international

---

- **Etats-Unis**
  - Campagnes d'évaluation NIST - DARPA
    - Depuis 1987, ouvertes en 1992, Ecrit / Oral
    - Production / distribution des ressources (LDC)
    - Organisation des campagnes (NIST)
  - Traitement du Langage Parlé
    - Reconnaissance, transcription, dialogue, reconnaissance locuteur...
  - Traitement du Langage Ecrit
    - Recherche d'information, compréhension, traduction automatique...
  - Traitement du Langage Ecrit + Parlé
  - Anglais américain, puis autres langues (TREC, TIDES)
  - Traitement de l'image (Reconnaissance de caractères / de visages...

# History of Speech Recognition Benchmark Tests



# Contexte international

---

- Japon
  - NII (extraction d 'information)
- Cocosda / Oriental Cocosda
- Conférences
  - HLT '02 (San Diego, 24-27 Mars 2002)
  - LREC '02 (Las Palmas, 27-31 Mai 2002)
  - LangTech '02 (Berlin, 26-27 Septembre 2002)

# Contexte européen

---

- 5ème PCRD
  - Human Language Technologies
  - Projets Elsnet, ELRA, Euromap, Enabler, Class, CLEF...
- 6ème PCRD (2003-2006)
  - Information Society Technologies (3,6 G€)
    - Knowledge Management & Multisensorial interfaces
  - Nouveaux instruments
    - Projets intégrés
    - Réseaux d 'excellence
    - Construction de l 'Espace Européen de la Recherche (ERA-NET)
    - Article 169, Réseau d 'infrastructure...

# Contexte national

---

- Actions de l'AUF (Francil, ARP, ARC)
- Actions de la DGLF-LF
- Actions du CNRS
  - Ingénierie de la langue
- Associations
  - GFII, GESTE, AFTEL...
  - ATALA, AFCP, APIL, Evaling...

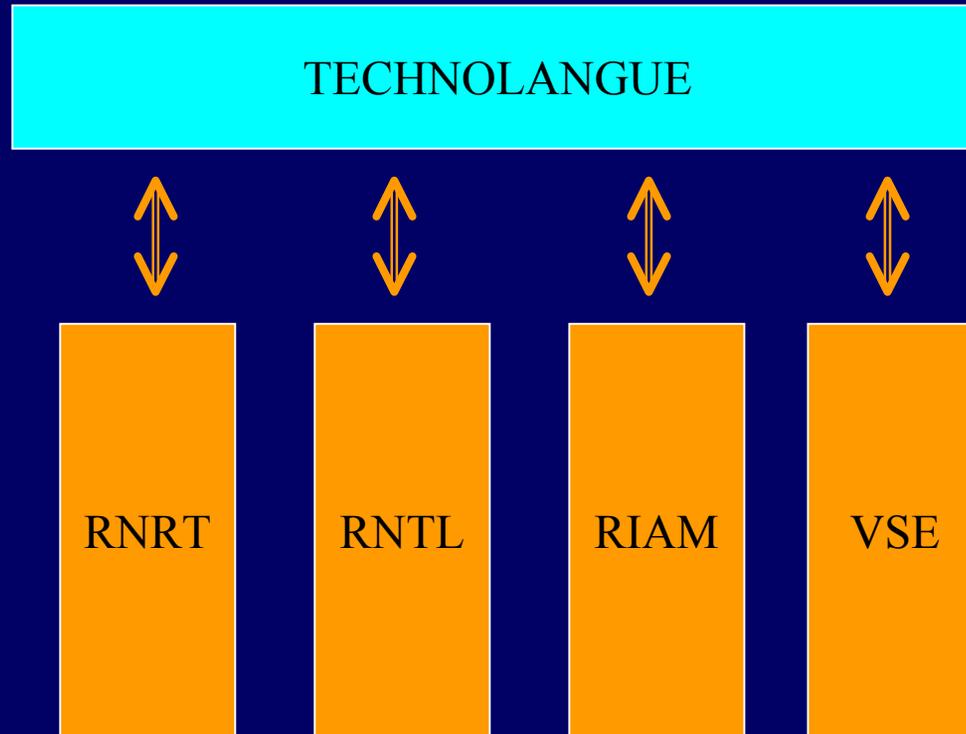
# RRIT

---

- Réseaux de Recherche et d'Innovation Technologiques (RRIT)
  - RNRT (Télécoms), RNTL (Logiciel), RIAM (Audiovisuel et Multimédia), RMNT (Micro-Nanotechnologies)
  - Projets de R&D de type exploratoire (MR) ou précompétitif (MinEFI, MCC)
  - Projets coopératifs (recherche publique - industrie)
  - Financé par le FRT (MR : 150 M€ / 38 M€ TIC)
  - Intérêt pour les technologies de la langue : Journée RNRT-RNTL “Langue et Technologies”, Nancy, Novembre 2000
- Veille Stratégique pour les Entreprises (VSE)
  - AAP 2002 « Veille sur Internet », 700 K€

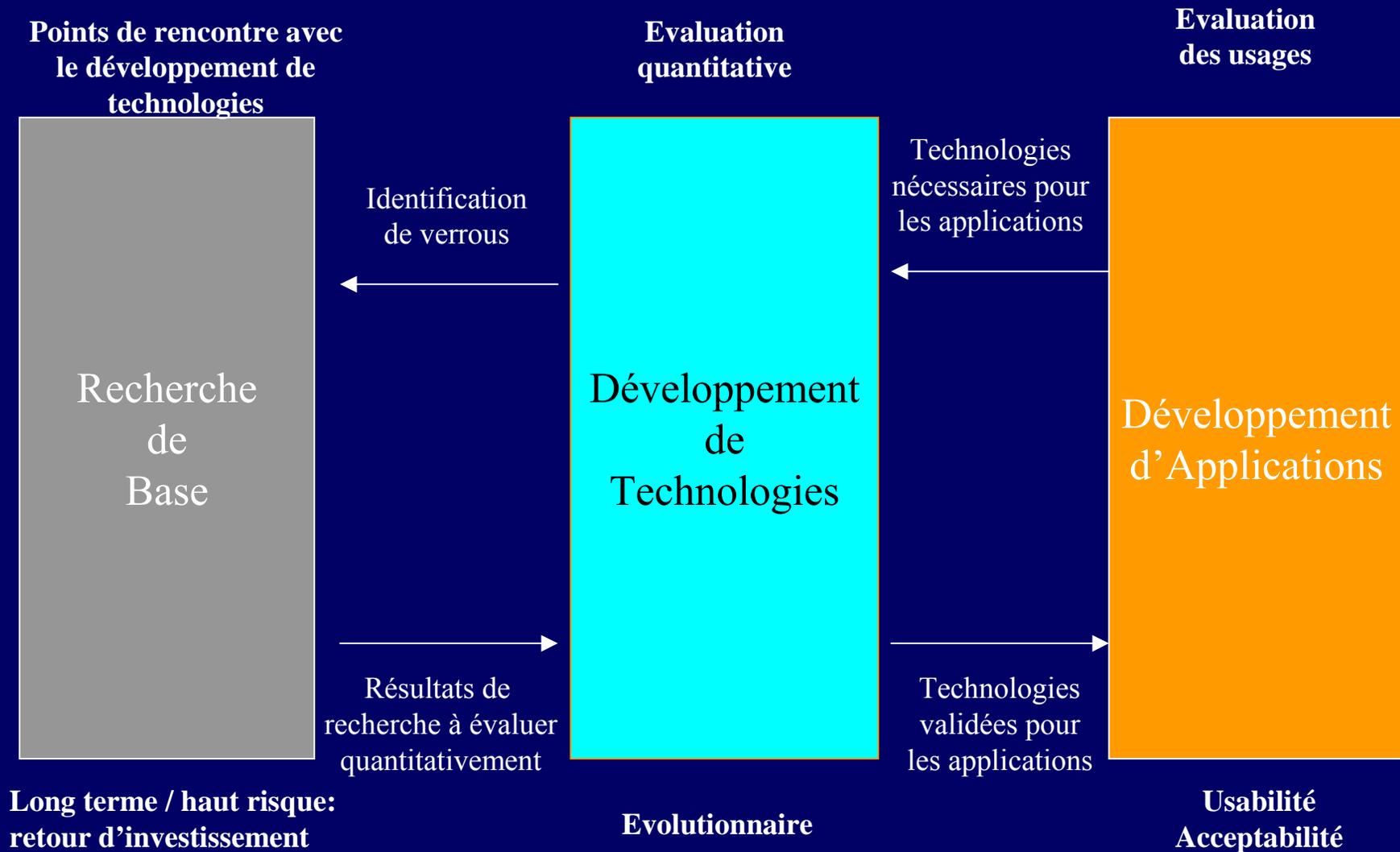
# Action « TechnoLangue »

---



Programme d'infrastructure en soutien à la R&D, la R&D restant dans les RRIT et le programme spécifique "Veille"

# Recherche Technologique de Base



# Action « TechnoLangue »

---

- **Accompagnée par des études**
  - Traduction automatique
  - Tissu industriel et recherche
- **Action similaire dans d'autres domaines**
  - Technologies de l'image...

# Action « TechnoLangue »

---

- **Organisation**

- Un Comité de Pilotage (CP) présidé par C. Fluhr (CEA)
- Composé de 15 personnes :
  - Représentants des RRIT: B. Bachimont (INA - RIAM), C. Sedogbo (Thalès - RNTL), C. Waast (IBM - RNRT)
  - Recherche publique: C. Fluhr (CEA), E. Geoffrois (DGA) P. Paroubek (Limsi-CNRS)
  - Industriels: K. Choukri (ELDA), B. Normier (Lingway), J.-J. Rigoni (Elan Informatique ), F. Segond (Xerox) + C. Sorin (FT R&D)
  - Administrations: S. Chaudiron (MR), J. Mariani (MR), D. Malbert (MCC), J. Mathieu (MinEFI)
- Equilibres recherche/industrie - écrit/parlé

# AAP « TechnoLangue »

---

- **Objectif : mettre en place un dispositif pérenne pour:**
  - Stimuler la production, la validation et la diffusion de ressources linguistiques
  - Organiser l'évaluation des technologies de la langue
  - Soutenir la participation des acteurs français aux instances de normalisation et de standardisation
  - Mettre en place un dispositif de veille informationnelle

# AAP « TechnoLangue »

---

- **Afin de :**
  - Encourager l'accès des centres de recherche et de formation aux ressources produites
  - Développer des compétences dans les technologies de la langue
  - Favoriser leur appropriation de manière large
    - Diffusion vers des filières de formation (Universités, IUT, Ecoles d'ingénieurs...)
    - En assurant un retour d'expérience
  - Renforcer les activités autour de la langue française et de son usage dans tous les domaines

# AAP « TechnoLangue »

---

- 4 réunions du Comité de Pilotage (janvier - mars 2002)
- Un appel à propositions en 2002 sur les 4 volets identifiés
  - Volet 1: Ressources linguistiques
  - Volet 2: Évaluation
  - Volet 3: Normes et standards
  - Volet 4: Veille informationnelle
- Un contenu de l'AAP rédigé pour chacun des 4 volets
- Consultation auprès des 3 RRIT TIC

# AAP « TechnoLangue »

---

- Ouverture internationale
  - Programmes nationaux voisins
    - Pays de l'UE
      - Allemagne, Italie, Norvège, Espagne
      - Grèce, Pays-Bas...
    - USA, Japon, Afrique du Sud...
  - Coopération avec partenaires étrangers
    - peuvent participer au projet
    - contribution sera rédigée en français
    - financement sur leurs fonds propres

# Volet 1 - Ressources

---

- Stimuler la production et la diffusion de ressources linguistiques pour :
  - répondre aux besoins minimaux (*Basic Language Resource Kit*) pour la langue française ;
  - favoriser la réutilisabilité des ressources ;
  - soutenir la recherche ;
  - aider au développement d 'applications industrielles ;
  - diminuer le coût du « ticket d 'entrée » dans le secteur.
- Nécessairement pour la langue française et, éventuellement, en association avec des langues partenaires.

# Volet 1 - Ressources

---

- **Données linguistiques orales et écrites :**
  - corpus oraux, lexiques de prononciation, etc.
  - bases de données pour la synthèse de la parole ;
  - corpus textuels mono et multilingues (parallèles, alignés...) ;
  - lexiques, terminologies, grammaires,...
  - ressources lexicales sémantiques : ontologies, thésaurus,...
  - corpus multimodaux,...
- **Outils logiciels « de base » :**
  - analyseurs morphologiques, syntaxiques, sémantiques,
  - extracteurs de terminologie,
  - identificateurs de langue,
  - outils d'étiquetage de corpus,
  - lemmatiseurs,...

# Volet 2 - Evaluation

---

- Organiser l'évaluation des technologies de la langue
  - Évaluation de technologies
    - Organisation de campagnes d'évaluation sur des secteurs non couverts par des programmes européens ou internationaux (cf TREC, CLEF...) ou selon des méthodologies différentes.
  - Évaluation d'applications
    - Développer des méthodologies d'évaluation pour des produits industriels ou pré-industriels, pouvant aboutir à des «boîtes à outils».
  - Réflexion sur les méthodologies d'évaluation
    - Améliorer les méthodologies d'évaluation existantes. Identifier de nouvelles approches (usabilité, aspects cognitifs). Nouvelles applications (Réalité virtuelle, interaction mulimodale, Internet...)

# Volet 3 - Normes et standards

---

- Soutenir la participation des acteurs français aux instances de normalisation et de standardisation
  - Présence très faible des acteurs français dans les instances de normalisation et de standardisation ;
  - Importance stratégique d'anticiper sur les normes et les standards industriels et de participer à leur élaboration ;
  - Multiplicité des lieux où s'effectue le processus de normalisation : comités officiels ou non, forums, projets,...
  - Soutenir la création de consortiums pour renforcer la présence française dans les instances (JTC1, CEN, W3C,...) ;
  - Aider la mutualisation des efforts sur chacun des thèmes concernés;

# Volet 4 - Veille informationnelle

---

- **Mettre en place un dispositif de veille informationnelle**
  - Créer un portail sur l'ingénierie linguistique pour permettre l'accès à l'information :
    - le panorama de l'offre industrielle et technologique,
    - état de l'art scientifique et technologique,
    - normes et standards,
    - appels à propositions...
  - En particulier pour les « nouveaux entrants »
  - En s'articulant avec les sites et les projets existants

# Procédures

---

- Calendrier
  - 10 juin : clôture de l'appel et envoi de la version électronique ;
  - 17 juin : envoi de la version papier (documents signés) en deux exemplaires et accusé de réception ;
  - 26 juillet : annonce des résultats.

# Procédures

---

- Procédure de sélection
  - Les dossiers incomplets ne sont pas pris en compte ;
  - Tous les projets doivent être expertisés pour être sélectionnés ;
  - Au cours de l'instruction, les porteurs de projets peuvent être contactés par le Secrétariat du Programme ou les rapporteurs du CP.

# Procédures

---

- **Procédure de sélection**

- Chaque projet est évalué par deux experts (recherche / industrie) soumis à une stricte confidentialité ;
- Avis motivé écrit envoyé au Comité de Pilotage (CP) par chaque expert ;
- Chaque projet est suivi par un membre du CP (rapporteur) et un chargé de mission ;
- Processus d'évaluation : les personnes du CP impliquées dans un projet ne participent pas à son évaluation ;
- Classement des dossiers en A / B / C
- Information auprès des 3 RRIT TIC

# Procédures

---

- **Procédure de financement**

- Chaque projet est orienté vers un des ministères par le CP ;
- Chaque ministère finance en fonction de ses procédures ;

=> le Ministère de la jeunesse, de l'éducation nationale et de la recherche

=> le Ministère de l'économie, des finances et de l'industrie

=> le Ministère de la culture et de la communication

# Technolangue

---

*Les premiers résultats...*

# Technolangue

---

- *Quelques chiffres :*
  - Projets reçus : **52** (mais **50** projets évalués car 1 inéligible et 1 retiré par les porteurs)
  - Total des coûts des projets : **35,9 M€**
  - Total des demandes de subvention : **21,7 M€**

# Technologies

---

# Technolangue

---

## Volet 1 : Ressources linguistiques et logicielles

- Cluster « Données écrites » : 19 projets (11,4 M€ / 7,4 M€)
- Cluster « Données orales » : 3 projets (3 M€ / 1,9 M€)
- Cluster « Outils de base » : 10 projets (10,2 M€ / 5,9 M€)

# Technolangue

---

## Volet 2 : Évaluation

- Cluster « Technologies de l'écrit » : 5 projets (1,6 M€ / 0,9 M€)
- Cluster « Technologies de l'oral » : 5 projets (2,9 M€ / 1,7 M€)
- Cluster « Méthodologies » : 2 projets (1 M€ / 0,6 M€)

==> pas de projet sur le thème « Évaluation d'applications »

# Technolangue

---

## Volet 3 : Normes et standards

- Cluster « Technologies de l'écrit » : 4 projets (1,9 M€ / 1,2 M€)
- Cluster « Technologies de l'oral » : 2 projets (2,1 M€ / 1,1 M€)

# Technolangue

---

## Volet 4 : Veille

- 2 projets reçus (1,4 M€ / 0,7 M€)

# Technolangue

---

## *Premiers commentaires...*

- Toutes les thématiques n'ont pas été couvertes (évaluation d'applications) ;
- Les thématiques ont été couvertes très inégalement (30 projets sur l'écrit, 10 projets sur l'oral, la majorité des projets « Outils de base » concernent l'écrit) ;
- Grande hétérogénéité des clusters « Ressources » en terme de couverture linguistique (langue générale/langues de spécialités).

# Technolangue

---

## *La suite du processus...*

- Retour d'expertises en cours (2 expertises (recherche publique/industrie) par projet) ;
- Prochaine réunion du CP le 16 juillet ;
- Annonce des résultats vers le 26 juillet ;
- Début des négociations financières en septembre (avec possibilité de recomposer les consortiums).

-----

# AAP « TechnoLangue »

---

- Adresses

- Mél : [technolangue@technologie.gouv.fr](mailto:technolangue@technologie.gouv.fr)

- Poste :

- Ministère de la Jeunesse, de l'Éducation Nationale  
et de la Recherche

- Direction de la Technologie

- Programme Technolangue

- 1, rue Descartes

- 75231 PARIS Cedex 05

<http://www.recherche.gouv.fr/appel/2002/technolangue.htm>