Jocelyn Pierre

La langue au cœur du numérique

Délégation générale à la langue française et aux langues de France

Les enjeux culturels des technologies de la langue

Février 2007

La langue au cœur du numérique

Les enjeux culturels des technologies de la langue

Jocelyn Pierre, ingénieure de recherche
du ministère de la Culture et de la
Communication

Février 2007

Préface

Les technologies de la langue sont une composante majeure de la société de l'information. Qu'elles soient disponibles sur le marché ou en cours de développement, ces technologies offrent des ressources inestimables pour tous ceux qui ont à produire, transformer, rechercher ou comprendre des « données ». Outils d'analyse et de connaissance de la langue, de reconnaissance et de synthèse vocales, moteurs de recherche sémantiques, logiciels de traduction… sont appelés à modifier en profondeur nos modes d'agir professionnels, comme nos actes les plus quotidiens. Leur développement n'est pas sans conséquences sur les politiques de la langue.

Le rapport établi par Jocelyn Pierre dresse un panorama détaillé des outils et des dispositifs de soutien existants, aux plans national comme international. Mais il va plus loin, car il propose un cadre de nature à mieux coordonner l'action publique. En effet, il est indispensable aujourd'hui d'améliorer la recherche dans ce secteur, de lui trouver de nouveaux débouchés et de favoriser l'appropriation des outils disponibles par le plus large public. Le rapport met aussi en évidence la responsabilité particulière des institutions européennes, qui doivent s'impliquer davantage dans les activités de recherche et de développement.

Parce qu'elles transforment et facilitent l'usage comme la diffusion des langues, qu'elles contribuent à modifier les rapports de force et d'influence qui existent entre elles, les technologies de la langue intéressent très directement le ministère de la Culture et de la Communication, garant de la politique de notre pays en faveur de la diversité culturelle et linguistique.

Dans un secteur où nos entreprises sont particulièrement performantes et créatrices d'emplois, il revient en effet à notre ministère - aux côtés des différentes institutions et administrations concernées - de mettre en évidence à quel point les industries de la langue servent le renforcement de l'usage du français dans la vie sociale et la promotion de la diversité linguistique dans les situations de communication internationale.

Xavier North
Délégué général à la langue française
et aux langues de France

Sommaire

Enjeux et propositions

« Where is the wisdom we have lost in knowledge ?
*Where is the knowledge we have lost in information ? »**
T.S.Eliot, « The Rock », 1934

Malgré l'immensité et l'évidence des enjeux dans tous les secteurs, à l'heure où le numérique devient omniprésent et où la langue devient centrale dans les processus informatiques, mais où la recherche sur les technologies et sur les usages est dans le creux de la vague, la prise en charge institutionnelle du dossier se cherche. Sans accepter la fatalité et sans nier la réalité, l'objectif est de réagir.

La rencontre de l'informatique et de la linguistique a, depuis près de 50 ans, donné lieu à de nombreuses recherches, développements technologiques et intégration dans des services applicatifs. Les objectifs traditionnels recherchés étaient la traduction, l'interrogation de bases de données en langage naturel, le dialogue optimal entre l'homme et la machine, la reconnaissance et la synthèse de la parole. Depuis environ 10 ans, la croissance de l'internet et sa convergence avec les autres médias devenus numériques (le téléphone, la télévision, la radio, la presse) ont suscité l'apparition de nouveaux besoins. Des techniques de traitement des données textuelles ont fait leur apparition pour sélectionner, classer, structurer, etc. l'ensemble des informations disponibles sous forme numérique, qu'elle soient écrites, orales ou multimédias, monolingues ou multilingues.

Ce document a vocation à offrir un diagnostic d'actualité sur les technologies de la langue (chapitre 3) et le paysage institutionnel dans lequel s'inscrivent les activités économiques qui les génèrent (chapitre 4) ; à lancer quelques pistes d'interrogation sur l'actualité des missions du ministère de la Culture et de la Communication au vu du développement actuel de ces technologies (chapitre 1) et à envisager des pistes d'actions afin que leur développement soit utilement soutenu par les pouvoirs publics, le ministère de la Culture et de la Communication en particulier (chapitre 2).

---

* Où est la sagesse que nous avons perdue avec la connaissance ? Où est la connaissance que vous avons perdue avec l'information ?

Enjeux
et propositions

La langue au cœur du numérique

Les nouvelles tâches dévolues aux ordinateurs, de plus en plus complexes, ont une dimension langagière croissante au détriment des simples fonctions de calcul. Compte tenu du triple rôle de mémorisation, de cognition et d'échange de la langue, **au fur et à mesure du développement de l'informatique, le langage humain tend à y prendre une place de plus en plus centrale et essentielle.** Le mouvement ira en s'accélérant puisque l'objectif de notre système économique et politique actuel, fondamentalement technologique, est de repousser la limite de la division du travail entre humains et ordinateurs.

Le numérique tend à devenir le dénominateur commun de notre société dite « de l'information ». L'information numérique, textuelle (écrite ou orale, plurilingue) ou codée (parce que sonore ou visuelle), est désormais massivement présente et au cœur des activités économiques et sociales. La croissance exponentielle des informations disponibles sur les différents supports, publics ou privés, (télévision, radio, internet, bases de données, archives) est combinée à leur hétérogénéité (texte, voix, image et vidéo). Elle provient du passage de la production des contenus des méthodes analogiques aux méthodes numériques, de la **numérisation** des contenus existants et de la **convergence,** c'est-à-dire l'addition et la combinaison, de l'ensemble des réseaux (et donc des secteurs d'activités afférents) jusqu'ici séparés. Elle provient aussi de la libre mise à disposition d'outils logiciels permettant à l'ensemble des utilisateurs de devenir des producteurs de contenus.

Dès lors, intégrées à une multitude d'applications qui rendent des services tels que la veille informationnelle, la gestion documentaire, la traduction, la formation, la communication homme/machine, etc. les **« technologies de la langue » sont massivement et de façon croissante, sollicitées** pour assurer les tâches complexes que sont la production, l'accès, l'analyse, la transformation et la diffusion des contenus numériques multimédias et multilingues présents sur les réseaux. Le développement des « technologies du langage humain » pour reprendre le terme anglais[1] comme la régulation des effets de ce développement passe inexorablement par l'invention d'une nouvelle grammaire numérique. Cette nécessaire invention invite à **un double recentrage de la politique culturelle autour de la place centrale de la langue** dans le traitement informatique des contenus (1re partie) **et de la politique (multi)linguistique par le biais de l'outillage informatique** de la langue (2e partie).

[1] *Human language technologies* ou HLT.

La langue au cœur du numérique culturel

Pour comprendre l'impact de la vague numérique sur le secteur culturel et sur la politique culturelle, il est possible d'avoir en tête les débats de ces dernières années sur le droit de la propriété littéraire et artistique. Le droit a été métamorphosé dans sa structure (les notions d'œuvre et d'auteur sur lesquelles repose ce droit doivent être ré-interrogées), dans son « assiette », car ce droit entame son rayonnement vers des secteurs plus vastes et plus rémunérateurs. Longtemps « cantonnée », l'utilisation de moins en moins « exceptionnelle » de ce droit culturel montre à quel point **c'est toute l'autonomie générale du secteur créatif, patrimonial, symbolique… qui s'étiole.** Or, **la prise en charge de ces « contenus » par les institutions culturelles** va bien plus loin que le débat règlementaire.

La combinaison entre l'informatique et les ressources linguistiques est au cœur de la transformation des technologies de l'information en technologies cognitives. Les « contenus » numériques, mais aussi bien sûr langagiers, s'intègrent à l'ensemble de nos références culturelles. Les méthodes pour les produire, les diffuser et y accéder organisent industriellement, mais aussi artisanalement, de nouvelles manières d'échanger, de lire, d'écrire, de se souvenir, de raisonner, d'imaginer. **Dans cette re-codification générale de l'information et du savoir, l'informatique apparaît comme l'élément dominant parce qu'elle est l'élément nouveau, mais elle ne doit pas occulter le terreau plus essentiel encore qu'est la langue. Pour le moment, la partie linguistique de cette révolution culturelle est encore un « trou noir » impensé.** Une paresse intellectuelle s'est installée, repoussant loin sur l'agenda politique la question des conséquences culturelles, encouragée par les promesses non tenues par les technologies de l'ingénierie linguistique trop souvent regardées comme, et présumées être, des outils culturellement neutres. La réflexion autour de cet aspect cognitif et langagier de l'outil permettra de sortir de la perspective uniquement économique, oublieuse de l'homme, dans laquelle s'inscrit aujourd'hui la construction de la « société de l'information » comme un mariage *a minima* entre des tuyaux et des industries culturelles.

En s'attachant aux **grandes missions traditionnelles du ministère de la Culture et de la Communication que sont le patrimoine, la création et l'accès du plus grand nombre aux œuvres de l'esprit,** il est nécessaire de remettre la langue au centre des grandes problématiques culturelles numériques actuelles. Parce que la langue sert à la fois à accéder à l'information, à réfléchir, à mémoriser et à communiquer.

De l'industrie de la langue à l'industrie de la connaissance

Les patrimoines sous forme numérique, contenus stockés sur différents supports, constituent notre mémoire. Les traces de la vie quotidienne, administrative, professionnelle, les œuvres d'art, la littérature scientifique, les archives audiovisuelles, la sédimentation des pages web, etc. témoignent de l'évolution de notre société. La question de la constitution, de la préservation, de l'organisation et de l'accès aux patrimoines est cruciale, pour élabo-

rer notre histoire et notre vision géopolitique du monde, pour préserver la diversité culturelle, pour alimenter l'économie de nos industries culturelles. Néanmoins, **l'abondance, loin de rendre l'environnement plus intelligible, peut en éloigner le sens.** Se présente dès lors le défi de l'exploitation de quantités massives d'informations dont les outils intègrent en leur cœur des technologies de la langue. **L'enjeu, c'est l'accès aux contenus, y compris aux contenus non sémantiques que sont les images, les gestes, les sons, etc.**

Tactiquement, il s'agit de replacer le sujet de l'industrie de la langue dans une problématique plus vaste, plus stratégique, car mieux connue et mieux reconnue politiquement, qui est celle des **« industries de la connaissance ».** L'objectif concret étant de disposer d'outils de « navigation » performants, multimédias et plurilingues rendant possible l'interrogation de la totalité d'une base documentaire ou du web à partir d'une seule requête dans une seule langue et en langage naturel. Dans l'espace numérique, **disposer de moteurs de recherche performants et fréquentés** est un enjeu industriel. C'est également une question de souveraineté, d'indépendance et de respect du pluralisme. Les outils informatiques d'accès aux contenus, au tri, à la proposition de réponse ne sont pas neutres. Ils reposent sur des choix qui révèlent les valeurs et les normes de pensée de ceux qui les ont conçus.

Les nouvelles formes de l'intelligence collective

« Quelqu'un va-t-il prendre enfin la défense de l'infini ? »
Louis Aragon, 1927

Le web 2.0 peut être considéré comme la figure exemplaire des « nouveaux usages » des médias par lesquels l'usager devient à la fois consommateur, producteur, éditeur et diffuseur : un mouvement social d'appropriation des outils en réseau, opposé à une consommation « passive ». Il ne s'agit ni d'un nouveau modèle économique ni d'une révolution technologique, mais bien de nouveaux usages d'outils qui permettent la mise en commun de la production intellectuelle de communautés d'utilisateurs. Le terme de web 2.0 souvent utilisé pour désigner ce qui est perçu comme une transition importante du web, passant d'une collection de sites statiques ou dynamiques à une plate-forme informatique à part entière, fournissant des applications web aux utilisateurs. Le système de la critique des livres par les clients sur Amazon.com ou l'encyclopédie Wikipédia sont des exemples simplistes, mais éclairants, de cette prise de parole multiple qu'il faut imaginer sur le contenu infini du web. Au cœur de ce web 2.0 se trouve le concept de folksonomie[2]. **La dimension linguistique de ces nouveaux usages est essentielle et double : cognitive** du fait des procédés d'indexation et de la création d'ontologies[3], et **plurilingue** parce que chacun doit pouvoir accéder à tout et être lu de tous.

---

[2] La définition du grand dictionnaire terminologique québécois est : Système de classification collaborative et spontanée de contenus internet, basé sur l'attribution de mots-clefs librement choisis par des utilisateurs non spécialistes, qui favorise le partage de ressources et permet d'améliorer la recherche d'information. Les « mots-clefs » sont généralement appelés tags ou étiquettes en français.

[3] Employé ici dans son sens informatique dont le grand dictionnaire terminologique québécois donne le sens suivant : ensemble d'informations dans lequel sont définis les concepts utilisés dans un langage donné et qui décrit les relations logiques qu'ils entretiennent entre eux.

Comment les pouvoirs publics peuvent-ils bénéficier de ces nouvelles formes d'intelligence collective[4] et non pas être dépassés par elles, jouant ainsi leur rôle de relais vers les citoyens ?

Parce que les citoyens n'ont pas accès aux grands médias, un nombre croissant d'entre eux se tournera vers ces outils, parallèlement à leur simplification et leur diffusion sur d'autres supports que l'ordinateur. Le premier rôle des pouvoirs publics est probablement d'aider chacun à s'approprier ces objets, d'accompagner des communautés de partage et d'usage conformes aux valeurs de la République (la langue, la propriété intellectuelle, le respect des données personnelles, les interdits sociaux, etc.) à travers l'école, mais aussi les bibliothèques, les médias audiovisuels publics, les sites internet publics dans une logique éducative, mais aussi une logique d'accompagnement des pratiques culturelles « amateurs ». Ceci est particulièrement vrai pour ce qui concerne les catégories de la population, qui pour des raisons d'éducation ou d'âge, n'ont pas été préalablement formées à la « lecture savante » et au rationalisme cartésien. Il faut réinventer et **ré-enseigner la « lecture » à l'heure de la navigation, des liens hypertextes, du multimédia, des contenus multilingues qui sont les éléments de la nouvelle complexité du monde. La dimension linguistique** de cette réinvention est évidente et **a trois dimensions : le plurilinguisme, le sémantique, le cognitif.**

D'autre part, **les outils de veille et d'intelligence collective faisant remonter une information complexe qu'ils permettent aussi de valider, sélectionner et diffuser,** ce mouvement met fin à la rareté artificielle imposée par les éditeurs / prescripteurs traditionnels. Dès lors, le second rôle des pouvoirs publics en général, du ministère de la Culture et de la Communication en particulier pour son secteur, est probablement de **réinventer de la rareté,** des filtres de sélection (dénicher le Mozart dans la multitude des gamins « sampleurs ») afin de recréer une échelle du « beau et du bien », un consentement à payer la création originale, un service public (de la musique, des arts plastiques…), etc. Parce que le génie n'est plus rare, mais seulement imprévisible, **les experts publics,** tels que les conservateurs de musées ou de bibliothèques, **doivent développer un processus de validation *a posteriori,*** et non plus *a priori.*

La dimension culturelle de l'ingénierie linguistique

La langue, les langues, étant au cœur des missions culturelles de l'État, il s'agit pour lui de redéfinir en parallèle **une diplomatie linguistique au travers de l'outillage de la langue et du plurilinguisme.** Même si la « défense » de la langue française paraît aujourd'hui à cer-

---

[4] Le terme « d'intelligence collective » peut apparaître comme un anglicisme. Il est néanmoins largement utilisé par les professionnels et les penseurs de l'internet. *Stricto sensu,* l'intelligence collective est un concept régulateur qui peut être défini comme une intelligence variée, partout distribuée, sans cesse valorisée, coordonnée en temps réel, qui aboutit à une mobilisation effective des compétences. L'intelligence collective constitue également un champ de recherche dont l'objet est l'étude de la coopération intellectuelle entre humains dans un environnement techniquement augmenté. Enfin, l'intelligence collective est un projet « politique » dont l'enjeu est d'améliorer de manière notable les processus de collaboration intellectuelle au sein des réseaux de recherche, des entreprises, des administrations, des associations et des communautés virtuelles de tous ordres.

tains un acte inutile, voire déplacé à l'heure où se prône un plurilinguisme auquel on l'oppose paradoxalement, cet outillage informatique du français est le passage obligé pour qu'il reste une langue vivante. L'enjeu est double :

> Le français permet de produire, d'accéder à et de classer des contenus numériques quels que soient leur langue, leur format et l'interface de communication avec la machine (oral, etc.)

> Le français permet de communiquer avec l'ensemble des autres langues sur un mode automatisé (aujourd'hui l'anglais, demain les langues de tous nos voisins européens ainsi que le mandarin, l'arabe, le japonais, etc.).

Une nouvelle grammaire numérique

Le langage est au cœur de systèmes d'information de plus en plus nombreux et complexes. Il est donc essentiel de continuer « d'outiller », **« d'équiper » la langue française** afin qu'elle puisse continuer à être une langue de communication internationale, de travail, de communication avec les machines et non pas seulement une langue savante, de l'intimité ou du loisir. **Le défi technologique est de faire en sorte que la faculté de langage,** de notre langage, écrit et oral, **soit prolongée au sein des machines** à travers des besoins aussi divers qu'échanger, s'exprimer, créer, rechercher, classer, analyser, diffuser, reproduire, vérifier. C'est **une nouvelle grammaire numérique** de la langue qu'il faut continuer de concevoir. La langue doit être équipée pour dire le réel. Aujourd'hui, les langues sans dictionnaires ne sont plus des langues de travail ; demain, les langues sans appareillage informatique, sans outils performants de traitement automatique ne seront plus des langues de travail. La société de l'information est une société de mots, de thésaurus et d'index, qu'il faut tenir à jour, interroger, connecter et interconnecter... Par cette nouvelle grammaire, c'est aussi la régulation d'une certaine « normalisation » rampante de la langue qui trouvera un moyen d'expression.

Conséquemment et parallèlement, les « machines » doivent offrir un accès facile et convivial aux utilisateurs pour satisfaire les besoins d'un public sans formation linguistique ou informatique spécifique et doivent pour cela intégrer des solutions issues de l'ingénierie linguistique. Il est important de ne pas creuser le « fossé numérique » entre le « progrès » des outils logiciels intégrant des éléments linguistiques et la capacité du public à se les approprier. La « langue informatique », *a fortiori* si elle n'est pas une « langue française informatique », risque d'être vécue comme une langue étrangère par une partie de la population. Cette appropriation passe par un travail sur la terminologie d'une part et les interfaces d'autre part. **La langue est un puissant moyen de lever le barrage de « l'illectronisme ».** Afin de **développer des applications grand public effectivement appropriables par tous,** il est nécessaire d'intégrer des résultats d'ergonomie cognitive, d'anthropologie, de psychologie, et même de linguistique générale, etc. aux résultats de la recherche en linguistique computationnelle et en informatique. Or, ces résultats de recherche sont rares ou mal connus. La linguistique et l'informatique se sont renforcées en laissant tomber des pans entiers de recherche, y compris linguistique. Prenons comme exemple le succès des systèmes de consultation des sites internet par mots-clefs dans des pays non anglophones

comme la Chine, la Corée ou la Turquie qui demain seront des mots en langage naturel, des mots dictés et non dactylographiés… Insidieusement, les questions éternelles d'ambiguïté sémantique et de multilinguisme seront reposées. L'ingénierie linguistique pour le grand public est un phénomène social d'ampleur politique, culturelle.

Le plurilinguisme comme idéal technologique

« Unifier, c'est nouer mieux les diversités particulières, non les effacer par un ordre vain »
Saint-Exupéry, Citadelle, 1948

Les outils développés grâce aux technologies de la langue font naître le besoin et rendent toujours plus nécessaire une approche multilingue du monde. La « traduction » (non réduite à son acception strictement linguistique) apparaît sans aucun doute comme l'un des enjeux majeurs du multiculturalisme. Il faut pouvoir à la fois accéder et comprendre (être accédé et se faire comprendre) à l'immensité numérique du monde. Concernant la traduction automatique par exemple, il est très difficile d'inférer de l'existant les usages qui se développeront lorsque des outils pratiques, quasiment gratuits et proposant des résultats de meilleure qualité seront disponibles. Il est évident que s'en tenir à l'automatisation des besoins actuels est très en deçà de ce qu'il est nécessaire d'envisager. Dès lors, se représenter les enjeux devient un exercice difficile.

A minima, il y a lieu d'imaginer **les conséquences diverses de l'absence d'obstacles linguistiques** (vaincre la malédiction de Babel) **pour les textes écrits** (traduction et recherche d'une information même visuelle ou sonore) **et les échanges oraux** grâce à la combinaison des outils de traduction avec les outils de reconnaissance et de synthèse vocale. Un accès différent aux moyens de traduction fera évoluer le marché économique de la traduction, mais aussi, et surtout, évoluer le rôle social de celle-ci et donc la place des langues dans la société.

A maxima, il y a donc lieu de **repenser la place de la langue dans la société** (le point de fuite) : l'exclusion d'une majorité de la population du globe qui n'a pas accès au savoir global et à l'information, car elle est limitée à sa propre langue d'apprentissage ou est illettrée[5], la fragmentation du monde numérique en bassins linguistiques (déjà en mandarin, demain en hindi, en espagnol et en arabe) sous la pression de la communication numérique (l'internet, la télévision, les bases de données, le téléphone), etc. L'éternelle figure de l'« Hercule françois » qui tient d'une main une massue et de l'autre une chaîne d'or attachant à sa bouche les oreilles d'auditeurs captivés par ses paroles est plus que jamais d'actualité.

---

[5] Selon les Nations unies, la moitié des habitants de la planète ne sait ni lire ni écrire.

L'action publique : les atouts à jouer

« L'exercice du « multilatéral », c'est accepter dans l'espace numérique, d'appartenir à plusieurs cercles. Nous devons être, dans le même temps, français, européens, francophones et latins. »
Jacques Attali

Les 3 piliers traditionnels de l'action publique

Les enjeux ainsi décrits s'inscrivent bien évidemment dans les grands objectifs qui ont, depuis plusieurs décennies, encadrés toutes les actions prises en faveur des technologies de la langue, exprimés de façon implicite ou explicite : le rayonnement de la langue française et de la francophonie, la défense de la diversité linguistique et culturelle et le « patriotisme économique ».

La défense de la langue française est un devoir de l'État. Comme l'écrit clairement Hubert Astier dans son rapport [ASTIER, 2005] : *« La langue, maternelle, n'est pas un hochet de la nostalgie, elle est l'armature de toute solidité et donc de toute solidarité, nationales. Elle est, en outre, la matière innocente la plus riche par laquelle la diversité culturelle s'affirme au plus grand nombre. Car nous pensons comme nous parlons. [...] Il faut donc agir sans complexe. Une politique de la langue n'est pas un acte cocardier ou le dernier chant du cygne tant elle répond à un faisceau de raisons impérieuses, sociales, économiques et culturelles. [...] Il faut auparavant, se convaincre que la langue en est le vrai marqueur comme elle est aussi la principale arme (l'Histoire l'a toujours prouvé) pour refuser tout assujettissement inutile des nations, pour cause de facilités circulatoires, au plus petit commun dénominateur d'une Cité mondiale en gestation. »*

Citons là un autre rapport éclairant sur les responsabilités de la France francophone, celui de Patrick Bloche [BLOCHE, 1999] : *« La France a aujourd'hui des stratégies multiples parce qu'elle-même est multiple. La France est européenne, mais elle est aussi latine. [...] La France est francophone, mais elle est aussi méditerranéenne et il faut redoubler d'efforts pour arrimer le français aux langues de la Méditerranée, et en tout premier lieu à la langue arabe. La France entretient avec l'Afrique des relations parfois tumultueuses, mais fortes et chaleureuses, fondées sur la solidarité. La France entretient des liens étroits avec l'Amérique grâce au Québec, en premier lieu, mais aussi parce que l'histoire de France et l'histoire de toute l'Amérique ont toujours été alliées sur le socle de la démocratie. La France se porte aussi vers d'autres régions où elle est moins connue, avec lesquelles elle souhaite nouer des relations neuves d'échanges et de partenariats. »*

La diversité culturelle et la diversité linguistique sont un enjeu politique : y répondre est, pour la France, une condition essentielle de son influence. C'est cette idée générale qu'a reprise le ministre de la Culture et de la Communication, Renaud Donnedieu de Vabres, dans son discours lors de la Journée européenne des langues en septembre 2006. *« Car*

c'est bien de la diversité culturelle et de son expression la plus parlante, la diversité linguistique, qu'il s'agit ce soir, et cette manifestation s'inscrit très directement dans la politique que je mène dans tous les secteurs de la vie culturelle pour ***défendre le droit des peuples à la valorisation de leur patrimoine, à la libre expression de leur culture et de leurs créations.*** *Parce qu'au-delà de la France, bien au-delà même des frontières européennes, il s'agit d'une aspiration universelle. [...] La diversité culturelle ne se divise pas, et la diversité linguistique en fait partie intégrante. Mieux encore : parce qu'****une langue n'est jamais seulement un outil de communication, mais aussi un marqueur d'identité et un matériau de création,*** *la diversité des langues dans le monde en est l'expression privilégiée. »*

Paradoxalement, il est néanmoins nécessaire de noter que **la Convention internationale sur la protection et la promotion de la diversité des expressions culturelles de l'UNESCO[6] n'aborde pas la question de la diversité linguistique.** L'examen des débats semblerait montrer que nombreux sont ceux qui considèrent la problématique de la langue comme un « océan non maîtrisable » et préfèrent laisser cette question à l'écart de celle du respect de la diversité culturelle.

Enfin, l'argument économique est invoqué : ne laissons pas échapper définitivement ce marché de « l'accès » que Bill Gates définissait comme *« the next big thing »* il y a déjà 10 ans. C'est sur cette base que le Haut responsable chargé de l'intelligence économique au secrétariat général de la défense nationale a constitué de 2005 à 2006, un groupe de travail sur les outils de la traduction automatique. L'argument interventionniste est double et pourrait être résumé ainsi : le secteur de l'ingénierie linguistique et ses enjeux vont prendre une importance grandissante, pour l'État comme pour l'ensemble des secteurs économiques, dans la période actuelle de mondialisation de l'économie et de grandes mutations technologiques... **Pour l'État comme pour les entreprises, l'enjeu est double : lire les informations du monde dans la langue d'origine et attirer les étrangers sur les informations françaises (contenus et moteurs de recherche)** « afin de ne pas laisser aux Américains le monopole de la diffusion d'information sur le monde ». Second argument : **la France possède des technologies, des chercheurs et des sociétés reconnus ainsi que des industriels performants risquant d'être rachetés par des étrangers** qui récupèreront leur technologie alors que celle-ci a été largement soutenue par l'État français... Pour ces deux raisons ce secteur doit bénéficier d'un soutien politique : calculons l'investissement nécessaire à enclencher un effet de levier et intervenons. La démarche « d'intelligence économique » du SGDN est de valoriser ces enjeux afin de porter ces intérêts en lieu et place de lobbys inexistants. Le soutien à la recherche et à l'innovation permet de défendre la langue française et, à travers elle, de promouvoir les produits, les services et les normes français.

---

6 Adoptée le 20 octobre 2005 à la quasi-unanimité des 150 pays votants – moins deux voix contre, celles des États-Unis et d'Israël –, la Convention de l'UNESCO sur la diversité culturelle entrera en vigueur en mars 2007 car suffisamment d'États l'ont déjà ratifiée à ce jour.

Une politique publique en retrait

La puissance publique, nationale et communautaire, a investi de longue date ce dossier complexe des technologies de la langue sans arriver à lui donner la légitimité et la visibilité méritées, ni même un nom et une définition stables. La situation actuelle est un peu paradoxale. Nous disposons d'une part d'un corpus théorique de « philosophie politique » très conséquent autour des trois grands objectifs cités *supra* qui a donné lieu à de nombreux engagements internationaux. Nous disposons, d'autre part, de développements techniques très pointus en informatique linguistique. Mais l'entre-deux est plus délicat. Les applications techniques, directement utilisables, ne sont pas arrivées à « maturité ». **Il n'y a jamais eu de réelle politique publique pérenne développée pour articuler les technologies existantes ou en gestation et les principes théoriques.** Dans les années 1980, l'invention d'une étiquette « industrie de la langue » a permis de fédérer des activités existantes (la recherche en linguistique en particulier) et des activités innovantes (des applications informatiques). Cela a permis de médiatiser les enjeux et de drainer des fonds publics. Aujourd'hui, on rencontre une vraie difficulté à « cerner » le sujet et à stabiliser une appellation. **Le hiatus entre les enjeux, immenses, et les moyens mis en œuvre** afin de progresser significativement dans des domaines d'une infinie complexité (l'automatisation du langage humain) explique probablement ce déficit d'image et d'efficacité de l'action publique.

Les **politiques,** surtout internationales, ont été **plus incantatoires qu'actives.** La réflexion philosophique et politique autour du plurilinguisme, de la diversité linguistique, de la diversité culturelle, etc. en général puis dans la « société de l'information » est omniprésente et fort riche. Des tribunes comme l'UNESCO (l'Initiative B@bel par exemple), l'Organisation internationale de la Francophonie, le Sommet mondial société de l'information[7] et le Forum pour la gouvernance de l'internet[8], la « contribution pour une Europe numérique » présentée en juillet 2006 par la France à la DG Infso de la Commission européenne, en sont des exemples frappants. Malheureusement, la réflexion autour du « comment ? » et les programmes d'intervention afférents ne sont pas toujours à la hauteur. La multiplication des discours purement idéologiques, parfois sans pertinence économique, ainsi que la multiplication des « observatoires » tendant à parasiter les systèmes d'action sont probablement le signe d'un grand désarroi face à la question du comment.

Sauf exception, **le dossier a été porté à un niveau insuffisant.** Cela a débouché sur une **absence d'inscription pérenne dans le marbre des organigrammes administratifs :** ni

---

[7] Approuvé à Genève en décembre 2003, l'article 4 de la Déclaration de principes réaffirme l'énoncé de l'article 19 de la Déclaration universelle des droits de l'homme : « tout individu a droit à la liberté d'opinion et d'expression, [...] et celui de chercher, de recevoir et de répandre, sans considération de frontière, les informations et les idées par quelque moyen d'expression que ce soit... toute personne, où que ce soit dans le monde, devrait avoir la possibilité de participer à la société de l'information et nul ne devrait être privé des avantages qu'elle offre », la partie 8 de la Déclaration de principes et la partie C8 du Plan d'action sont consacrées spécifiquement à la diversité et à l'identité culturelles, la diversité linguistique et les contenus locaux.

[8] Cf. le symposium pour la promotion d'un internet multilingue tenu à Genève en mai 2006 sous l'égide de l'UIT et de l'UNESCO ou l'atelier sur la diversité linguistique tenu à Athènes en novembre 2006 dans le cadre du premier forum sur la gouvernance de l'internet.

direction ni directeur des technologies de la langue au ministère de la Culture, de l'Industrie, de la Recherche, à la Commission européenne, etc. Par ailleurs, devant sa complexité, peu de politiques ont choisi de faire de ce dossier un cheval de bataille national ou international. Par voie de conséquence, on peut constater deux effets collatéraux. Le premier est un **« manque de coordination entre les principaux ministères intéressés** et qui sont très normalement nombreux et donc une faible coordination générale des actions entreprises. » [ASTIER, 2005]. Le second est que **les nombreuses actions prises le sont rarement de façon continue,** pérenne, et que ces à-coups ont fait perdre beaucoup d'efficacité et de légitimité aux pouvoirs publics. L'absence de ligne spécifique aux technologies de la langue dans le septième PCRD comme dans les appels à projets de l'ANR pour 2007 en est l'actuelle illustration.

Sauf exception encore, **l'approche a été principalement technologiste.** Né **dans une dynamique de recherche,** le mouvement s'est trouvé peu à peu intégré à la grande dynamique « TIC ». À Paris comme à Bruxelles, les dossiers sont portés par les institutions Industrie et Recherche et, comme toutes les questions informatiques, par des ingénieurs (cf. par exemple, la représentation de la France auprès de la Commission européenne), trop peu sensibilisés ou sensibles à la problématique des usages, voire à l'anticipation du marché, et sont plutôt culturellement anglophiles. Il est ainsi sans cesse apparu comme essentiel de « travailler » la charnière entre les résultats de la recherche en informatique, en linguistique, en sciences de la documentation et les usages économiques, industriels, culturels... En 1995, un Conseil consultatif sur le traitement informatique du langage a été créé suite au rapport de André Danzin [DANZIN, 1995]. Dans la même idée, le rapport Bloche [BLOCHE, 1999] avait préconisé la création d'un Conseil des industries de la langue ce qui avait semblé contrarier les ministères de la Recherche et de l'Industrie. Le récent groupe de travail du secrétariat général de la défense nationale (SGDN) a lui aussi, rapidement senti la nécessité d'ouvrir son tour de table à des intervenants de différents horizons et aura eu ce mérite de rapprocher les acteurs et de contribuer à la visibilité politique du dossier.

Récemment, le secteur a eu tendance à se « privatiser ». L'arrivée de Google est exemplaire de ce changement. Les types de débouchés ont structurellement changé : pendant longtemps principalement militaires et, tout au moins très professionnels, **des débouchés commerciaux grand public apparaissent massivement** (moteurs de recherche internet, interfaces des appareils domestiques...). Les pouvoirs publics ont donc tendance à se retrancher dans une position qui privilégie la veille et la recherche fondamentale.

Le secteur des industries de la langue, à la fois secteur d'activités économiques et objet de l'action publique, est travaillé par deux grands mouvements qui touchent l'ensemble des secteurs. Ces deux grands mouvements mettent **en recul le niveau national / étatique des enjeux.** Comme ailleurs, on assiste **à la montée en puissance des niveaux *infra* et *supra* nationaux de l'intervention publique** parallèlement à la perte de substance du niveau national. Les ministères ont une action de plus en plus parcellaire et les moyens pour assurer des missions de coordination sont en constante diminution. Les directives, voire la règlementation, nationales ne sont pas toujours mises en œuvre par les ministères eux-

mêmes. Les financements se trouvent de plus en plus gérés au niveau des régions et de la Commission européenne. Comme ailleurs aussi, se font sentir les effets de **la « mondialisation » de l'économie.** Les grands groupes internationaux étendent leurs activités et leurs investissements dans le monde entier et cela rend encore plus vulnérables les PME françaises susceptibles d'être rachetées à vil prix du jour au lendemain. Les grands groupes français, investisseurs ou clients potentiels de PME nationales ou de produits francophones, ne considèrent plus l'enjeu national comme pertinent dans leur vision stratégique de développement. Il y a donc lieu de tenir compte de ce contexte global pour que les actions à mettre en œuvre gardent une quelconque utilité.

La langue : des solutions et non pas un problème...

... encore faut-il résoudre ses problèmes.

Devant ses enjeux, mais dans ce contexte, le ministère de la Culture et de la Communication a le **devoir** d'agir... dans les limites de ses moyens et de sa compétence. La **modestie** est de mise par manque de ressources humaines et budgétaires, mais aussi pour des raisons moins évidentes. En particulier, le sujet est complexe et le ministère ne dispose pas de l'expertise suffisante pour suivre sur le fond les dossiers de recherche et d'innovation de ce secteur. Parallèlement, ce dossier ne s'adresse pas aux « clients » habituels du ministère que sont les professionnels de la culture. Ce dossier a successivement été pris en charge, parfois partiellement, par différents services : la DGLFLF, le DAEI, la MRT. À ce jour, aucun emploi du ministère n'est affecté principalement au suivi institutionnel de ce secteur : l'absence du ministère dans les instances inter-ministérielles et communautaires en est la conséquence la plus directe. Néanmoins, pour le ministre de la Culture et de la Communication, réinvestir les politiques culturelles – notamment autour des pratiques numériques - et renouveler le discours sur la francophonie et le multilinguisme au travers de la question de l'équipement de la langue peut être un choix stratégique.

Créer la fonction

S'approprier le dossier des industries de la langue signifie pour le ministère, *a minima,* **la création d'un emploi budgétaire et le recrutement d'un chargé de mission de haut niveau** (type Haut fonctionnaire à...). La question du positionnement de cette personne est cruciale. La DGLFLF et la DDAI réunies englobent les missions et les compétences nécessaires pour alimenter l'instruction de ce dossier. Dès lors, tout positionnement qui ne serait pas véritablement transversal serait partiellement insatisfaisant. À défaut d'une délégation aux industries culturelles encore inexistante assurant une veille sur les industries de la connaissance (et donc aussi de la langue) et participant activement à la coordination *intra* et *inter* ministériel, **la DGLFLF (par l'entrée langue française) comme la DDAI, sont susceptibles d'assurer ce rôle.**

La DGLFLF peut s'appuyer sur son travail inestimable en matière de **terminologie,** ma-

tière première essentielle à l'alimentation de la recherche et des applications ; ses missions en matière de **plurilinguisme** et de traduction ; sa responsabilité en matière d'**utilisation du français dans les organisations internationales,** le secteur scientifique et le monde du travail - aucun de ces domaines n'ayant vocation à englober celui des technologies de la langue. Néanmoins, elle a perdu son statut interministériel et son implication dans la promotion des technologies de la langue ; elle est perçue aujourd'hui à l'extérieur principalement comme un organisme de défense de la langue française. Par ailleurs, **elle dispose de moyens trop faibles pour être plus qu'un observateur passif des évolutions du secteur.** Afin de participer activement à la conception des politiques et aux choix faits par les différents guichets, il est nécessaire de pouvoir contribuer aux financements.

La DDAI a dans ses missions le suivi des industries culturelles, de la recherche et des relations communautaires ; trois dimensions essentielles du dossier. Le DAEI a la charge de gérer la position ministérielle sur les questions communautaires, mais n'est pas positionné pour assurer des missions de niveau national, voire régional. Le DEPS est chargé des réflexions sur les usages culturels et l'économie du secteur culturel, mais n'a pas vocation à porter une position politique. La MRT est forte de son expérience en matière de numérisation et de ses compétences sur les problématiques de recherche. Néanmoins, la problématique de recherche seule laisse à l'écart les questions d'usages et d'innovation industrielle. Par ailleurs, à ce jour, la DDAI **n'assume pas de véritable coordination sur les industries culturelles,** laissant à la DDM et au CNC les principales actions sur le secteur (dont le suivi du RIAM), deux organismes qui, de fait, ne travaillent pas quotidiennement avec le minis-tère de la Culture et de la Communication.

Cet agent se verra confier **des missions formelles de coordination** avec la MRT et le DAEI de la DDAI, la direction du multimédia du CNC, le bureau Société de l'information de la DDM, l'ensemble des directions techniques puisqu'il y a lieu de rappeler que la problématique du numérique concerne aussi les dimensions patrimoniales du ministère, les directions régionales des affaires culturelles en particulier pour la prise en charge de la dimension innovation industrielle, les établissements publics via la DAG et de coordination interministérielle avec le SGAE, la direction de la langue française du MAE, le ministère de l'Industrie, le ministère de la Recherche, le SGDN, etc. Cette présence permettra de porter concrètement la problématique linguistique au cœur des politiques culturelles.

Militer pour un Technolangue 2

Un appel à propositions a été lancé le 17 avril 2002 sous le titre générique « Technolangue » regroupant les quatre volets fondamentaux de ce domaine : la constitution de ressources linguistiques, l'évaluation des performances des logiciels, le suivi de l'élaboration des normes et des standards ainsi que la veille technologique. Porté par les trois ministères chargés de l'Industrie, de la Recherche et de la Culture, financé par les trois réseaux nationaux de recherche : le réseau national de recherche en télécommunications (RNRT), le réseau national des technologies du logiciel (RNTL) et le réseau de recherche et innovation en

audiovisuel et multimédia (RIAM) pour un montant estimé à 6,2 M€. Le ministère de la Recherche a financé près de 80%. Tous les projets financés sur ce programme sont terminés.

Le caractère novateur et exemplaire du programme Technolangue est unanimement salué, notamment parce qu'il contient les éléments essentiels suivants : de la mutualisation des ressources, de l'évaluation, de l'appui à la normalisation, de la veille stratégique, des actions de promotion, surtout vers les industriels et une instance légitime de coordination. En outre, les experts du secteur semblent tous considérer que le programme Technolangue a atteint ses objectifs de départ.

Le programme Technolangue est arrêté depuis 2005. **Sa non-reconduction, y compris sous une autre forme, pose problème** et met les acteurs du secteur en position délicate. D'une part, les autres guichets semblent peu adaptés aux besoins du secteur ; soit ils sont plus tournés vers l'innovation et ne correspondent pas toujours aux contraintes des PME, ayant de très forts besoins en R&D avec des débouchés industriels incertains : les projets Eurêka, l'agence de l'innovation industrielle, les pôles de compétitivité, OSEO ; soit ils sont réservés aux chercheurs, mais le thème « traitement informatique de la langue » ne correspond à un axe prioritaire d'aucun des grands réseaux, RNTL, RNRT et RIAM. D'autre part, quelques projets financés sur diverses lignes ne font pas une politique et ne structurent pas un secteur. Enfin, l'absence d'une volonté politique nationale affichée fragilise grandement la position française dans les discussions communautaires relatives au financement de ce secteur.

Depuis la création de l'agence nationale de la recherche (ANR), les financements passent tous par ce guichet et **l'ANR n'a pas inscrit les technologies de la langue dans ses priorités,** même si quelques projets ont été ou sont acceptés au titre des appels à projets « masses de données et connaissances ambiantes » et « des données aux connaissances ». Les modes opératoires et les critères de choix de l'ANR restent encore flous pour les acteurs du secteur, y compris pour le ministère de la Culture et de la Communication, d'autant plus qu'ils sembleraient sur le point d'être modifiés en faveur de l'accroissement de l'influence du ministère de la Recherche. Pour continuer à bénéficier des effets positifs constatés du programme Technolangue, il est important de s'appliquer à la reconduction d'un programme équivalent, incluant probablement des éléments sur le multimédia et l'image, permettant de dégager de 3 à 4 M€ par an en interministériel de façon extrêmement visible.

Le dossier des technologies de la langue est pris au sérieux au ministère de la Recherche. Pour cela, il a besoin de s'intégrer à une dynamique gouvernementale avec le ministère de l'Industrie, le SGDN, le ministère de la Culture et de la Communication via la DGLFLF, le CNC pour le réseau RIAM, la MRT faisant valoir des besoins sectoriels (au même titre que la santé, les transports ou la défense) du ministère sur les bibliothèques numériques, les archives, l'indexation multimédia, etc. et non pas comme un laboratoire de recherche en quête de financement. Ainsi, **la DGLFLF et la MRT doivent ensemble instruire un dos-**

sier « technologies de la langue » pour le porter à l'ANR comme l'une des priorités du ministère qui contribuera à ce programme par des financements et par son expertise.

Enfin, lorsque l'ANR fait des appels à propositions pour la constitution de groupes d'experts, il est important d'envoyer des experts du secteur culturel, y compris de ses aspects linguistiques.

Dans l'intervalle, le ministère fera œuvre utile en contribuant à l'atteinte de certains des objectifs de Technolangue : la normalisation, la coordination du secteur et la constitution de ressources linguistiques en langue française.

S'appuyer sur une instance de coordination et d'impulsion

La constitution et le fonctionnement du groupe de travail piloté par le SGDN de 2005 à 2006 a montré à quel point **le sujet peine à trouver son étiquette, son périmètre, son modèle économique et son mode de régulation interministérielle.** D'abord appelé « intelligence économique », il s'est transformé en groupe de travail sur la traduction automatique. Constitué principalement de chefs d'entreprise, il s'est peu à peu étoffé de nombre d'administration montrant à quel point le secteur est dépendant des financements de la recherche et de l'innovation. Fort de la description des enjeux pour l'ensemble des secteurs, il n'a pas su déboucher sur des propositions concrètes et réalistes à la hauteur de ces enjeux ; même pas sur la reconduction du programme Technolangue. Ce constat est à la fois le signe de la crise d'un secteur peu structuré et de la difficile coordination interministérielle.

La question de l'outillage informatique de la langue **entre dans les missions de plusieurs ministères, mais, malheureusement, ne vise en priorité la « clientèle » traditionnelle d'aucun d'entre eux.** Le ministère de la Recherche est celui qui, jusqu'à aujourd'hui, a porté avec le plus de vigueur ce dossier bien qu'une bonne partie de la valeur ajoutée du secteur se trouve dans les sociétés privées. D'ailleurs, pour l'heure, seul **le ministère de la Recherche** dispose d'une personne chargée du dossier des industries de la langue, même s'il faut noter qu'elle relève de la mission d'information scientifique et technique. **Le ministère de l'Industrie** défend ce secteur auprès des guichets nationaux et communautaires, bien qu'aucun grand groupe n'y soit stratégiquement impliqué. Les enjeux du **ministère de la Défense** sont évidemment essentiels, mais ce ministère est rarement enclin à la coopération gouvernementale. Les **ministères de la Culture et des Affaires étrangères** se partagent la responsabilité de la « langue française », mais sans répondre à la demande de leurs clients habituels que sont le réseau diplomatique et les organisations internationales d'une part, les artistes et les grands corps patrimoniaux d'autre part. **Le SGAE** suit le dossier au niveau du Conseil des ministres et du Parlement, mais n'assure pas le rôle de coordination des experts techniques (des ministères de l'Industrie et de la Recherche) auprès des services de la Commission.

Renforcer la concertation interministérielle sur le traitement informatique du langage et ani-

mer un secteur en aidant ses acteurs à transcender leurs intérêts individuels peut se faire par deux moyens. Soit en créant un nouvel organisme, soit en assignant cette fonction à un organisme existant. Dans l'un et l'autre cas, **la création ou l'assignation d'une instance sera en soi un acte fort de légitimation du dossier par les pouvoirs publics.**

Une proposition est **la création d'une agence de la langue française** qui s'appuierait sur des exemples étrangers : la *Deutsh language union* (DLU) ou *Nederlandse Taal Unie* co-financée par les gouvernements belge-flamand, néerlandais et du Surinam ; la *Termcat catalane* financée par le gouvernement autonome de Catalogne et s'inscrivant dans une politique active[9] permettant à chaque citoyen de langue catalane de s'adresser à la Commission européenne dans sa langue ; le portail allemand LT World financé par le BMBF (ministère fédéral de l'Éducation et de la Recherche), etc. Il s'agirait de créer une grosse structure autonome juridiquement, comme une fondation, qui permettrait de répondre aux appels d'offre européens dans le domaine des industries de la langue. Cette hypothèse n'est envisageable qu'à partir d'une réelle volonté politique de financement. **Cette agence pourrait avoir vocation à devenir une agence européenne de la langue française et être co-financée par les autres pays européens francophones.** Elle regrouperait alors les trois fonctions essentielles que sont le financement (rôle que jouent actuellement les ministères de la Recherche et de l'Industrie, l'ANR ou la NSF américaine), la prescription (comme un conglomérat du type Technolangue), la mutualisation et la mise à disposition des ressources (comme ELRA ou la TST en Hollande).

L'autre solution, beaucoup plus modeste, est de **s'appuyer sur une instance théoriquement existante comme le Conseil supérieur de la langue française.** Ce dernier ayant vocation à être modifié à court terme[10], il pourrait se voir attribuer des missions relatives aux technologies de la langue qui seraient prises en charge par un groupe de travail *ad hoc* prenant modèle sur le Conseil consultatif sur le traitement informatique du langage, le comité de pilotage de Technolangue, ou une pérennisation du groupe de travail organisé par le SGDN. Ce réseau d'experts devra s'appuyer sur une équipe interministérielle, être visible et médiatisable, y compris au niveau communautaire, être sollicité par de grosses structures telles que le consortium Quaero ou le projet BNuE. À l'instar du Comité interministériel d'aménagement du territoire, cette instance de coordination se réunirait deux fois par an au moins et prendrait les décisions sur tous les sujets à portée interministérielle relatifs aux aspects intérieurs et extérieurs du secteur. **Après la nomination d'une personne qualifiée au ministère de la Culture et de la Communication, la prise en charge du secrétariat général de ce groupe pourra être un moyen d'exister dans ce secteur.**

Le groupe de travail ainsi constitué pourrait avoir **la mission de stimuler les débats natio-**

---

[9] Le gouvernement prenant à sa charge les frais de traduction de et vers le catalan pour les échanges communautaires, il y a là un terreau très fertile pour les PME « linguistiques » de Catalogne.

[10] Le décret du 2 juin 1989 instituant un Conseil supérieur de la langue française et une délégation générale à la langue française modifié par le décret du 21 mars 1996 et celui du 16 octobre 2001 doit faire l'objet de plusieurs modifications pour tenir compte d'évolutions intervenues récemment.

naux, de réaliser chaque année un état de l'art et de faire des propositions aux décideurs. Pour ce faire, de nombreux dispositifs d'observation, de comptabilité et de réflexion sont possibles : un séminaire de réflexion avec des intervenants extérieurs / étrangers, en particulier sur les conditions de l'accès au savoir numérique ; une mission de la Cour des comptes ou d'un inspecteur des finances sur les politiques menées dans le secteur ; une commission au Commissariat général au plan ; une ou des études économiques commandées au DEPS sur les marchés concernés (traduction, moteurs de recherche, veille informationnelle et stratégique, gestion documentaire), etc.

Mettre à disposition des ressources linguistiques

La mise à disposition de ressources linguistiques, qu'il s'agisse de corpus ou de ressources plus complexes comme des dictionnaires (la base de données du Trésor de la langue française sous format libre par exemple) est **une activité essentielle d'appui à la recherche.** Il s'agit d'un travail de fond, sans fin, incompatible avec un financement par les industriels à cause de son coût, de son calendrier et du profil des ressources humaines nécessaires (des chercheurs qualifiés). **N'ayant aucune justification économique sérieuse à court terme, cette activité incontournable ressortit à la responsabilité de l'État.**

Sur ce domaine, **le ministère de la Culture et de la Communication est légitime pour agir. La langue est sa responsabilité, en tout cas la langue française et les langues de France.** Le soutien à cette activité ne fait pas concurrence aux industriels, bien au contraire, si des modes de diffusion adaptés sont pensés (comme l'agence ELRA ou le projet de la DGLFLF sur les corpus oraux). Il n'obère pas non plus les missions des ministères de l'Industrie ou de la Recherche.

Cette action doit être intégrée à la politique générale du ministère relative à la mise à disposition des données publiques essentielles. Il s'agit d'une question de grande actualité puisque le rapport Lévy-Jouyet [LEVY-JOUYET, 2007] sur « l'économie de l'immatériel » insiste bien sur les enjeux liés à l'application de l'ordonnance n° 2005-650 du 6 juin 2005 relative à « la liberté d'accès aux documents administratifs et à la réutilisation des informations publiques » et que la directive communautaire de 2005 sur le même sujet vient d'être récemment transposée en droit français.

Appuyer la présence française dans les chantiers internationaux de normalisation

Le problème des accents dans le courrier électronique, encore imparfaitement résolu, a révélé aux non-spécialistes de ces questions **l'importance et les enjeux des normes et des standards pour les langues et les cultures.** Des normes techniques peuvent entraîner des conséquences majeures, par exemple sur la sécurité, la confidentialité, le multilinguisme ou la propriété intellectuelle, qui relèvent de questions sociétales. **Par ailleurs, la normalisation est essentielle, car elle est un lien naturel et important entre l'industrie, les chercheurs et les utilisateurs.** Elle permet de réaliser une veille technologique et une veille normative très efficaces et diffusables vers les industriels.

La normalisation française présente des difficultés conjoncturelles, à la mesure de notre faible présence dans les groupes de travail et les différents comités stratégiques. L'activité de veille culturelle et linguistique sur l'ensemble des champs de normalisation et de standardisation réalisée par la France est encore insuffisante. Cette carence est autant celle des industriels et des organismes publics concernés que des représentants des utilisateurs et des consommateurs, qui pourraient s'assurer que les normes élaborées favorisent l'intérêt général et non le seul intérêt industriel et commercial. La normalisation n'est pas toujours aussi centrale dans les réflexions qu'elle devrait l'être : le groupe de travail du SGDN sur « l'intelligence économique » n'a pas relevé ce secteur comme prioritaire par exemple.

La normalisation prend beaucoup de temps et doit être suivie par des professionnels accomplis. Probablement par manque de culture du lobbying, il n'y a pas de Français « professionnels » sectoriels de la normalisation dans les instances internationales. Les experts rencontrent trop fréquemment des problèmes de reconnaissance de cette activité par leur employeur, de décharges de travail et de frais de mission. L'AFNOR n'a pas de budget pour envoyer des missionnaires à l'ISO et de moins en moins d'argent pour recruter ses propres spécialistes. Les industriels français des technologies de la langue sont trop petits pour faire seuls face au coût de l'envoi d'experts, les quelques grands groupes industriels et les organismes publics sont trop absents.

Comme toujours, il serait nécessaire de réaliser / d'**actualiser un inventaire des acquis, des besoins et des ressources disponibles dans le domaine des industries de la langue** : les normes en vigueur, les travaux de normalisation en cours, les domaines où une participation plus active de la communauté industrielle et de recherche française serait nécessaire, les instances de normalisation *de jure* et *de facto* (près de 800), les experts francophones susceptibles de suivre les travaux de normalisation, les intervenants qui s'impliquent dans la défense des intérêts de la langue française et de son traitement informatique. Cet inventaire permettrait de déterminer des priorités et d'organiser une veille stratégique pérenne [DGLFLF / AFNOR, 2007 ; BOURBEAU, 1995]. La DGLFLF, l'AFNOR et l'OIF ont un rôle déterminant à jouer dans cette activité.

En matière de normalisation, la problématique de la langue peut s'entendre de deux façons : quant aux procédures et aux domaines normalisés. Sur le fond, **les deux mots de « plurilinguisme » et « d'internationalisation »** pourraient résumer les champs de bataille. L'objectif global est que la langue française, et à travers elle l'ensemble des langues, et l'ensemble des cultures puissent être prises en compte dans l'ensemble des applications informatiques, des langages, des interfaces. Relativement aux procédures, il faut noter que l'AFNOR a depuis toujours été exemplaire sur ce front en inscrivant comme une priorité la défense de langue française au sein du processus international de normalisation. Il est essentiel que cet activisme reste d'actualité et cela passe par la reconnaissance de **la langue française comme langue officielle de l'ISO,** actuellement en danger, par **la traduction systématique des normes et d'un maximum de documents de travail en français, par la constitution d'un réseau des instituts francophones de normalisation,** etc.

Cela passe aussi, évidemment et surtout, par la présence d'experts français et francophones dans les instances de normalisation.

Encourager l'implication active des Français et des francophones dans le processus international de standardisation est évidemment la pierre angulaire de l'action. La création d'un fonds d'aide à la présence française, au moins pour les personnes qui ont des responsabilités dans les instances de travail de l'ISO et des grands consortiums ne se fera pas sans difficultés. 22 000 experts français, issus du secteur public et pas assez souvent du secteur privé, sont mobilisés chaque année sur les chantiers de normalisation et ils sont généralement d'un très bon niveau : un fonds générique risque d'être submergé par les demandes. **Il semblerait plus adéquat de faire de l'appui occasionnel et ciblé, en fonctionnant par exemple par appel d'offres pour constituer des équipes de deux ou trois experts sur un thème.** Chaque administration pourrait alors contribuer spécifiquement selon ses besoins prioritaires. Les experts auraient un mandat clair et leur mission pourrait alors être évaluée. Parallèlement, il ne serait pas inutile de former les nouveaux entrants à la compréhension des enjeux globaux, des procédures, des structures, etc. de la normalisation et de réunir régulièrement et par thèmes les Français qui sont membres des groupes et des instances ISO. Parallèlement, un appui pourrait être apporté aux personnes morales françaises ou francophones, comme des associations professionnelles, acceptant d'être mandatées par l'ISO pour gérer les discussions autour d'une certaine norme (le Conseil international des archives ou l'IFLA le font déjà).

Enfin, quand les normes existent, il faut mener les actions nécessaires de promotion afin qu'elles soient mieux connues et que leur utilisation soit favorisée, notamment dans l'administration et les services publics. Un projet de l'ampleur de la bibliothèque numérique européenne constituerait également une vitrine exceptionnelle en matière de normes pour la diffusion, la description ou la conservation des données numériques.

Accompagner la prise en charge de l'aide à la recherche et à l'innovation

Les guichets d'aide aux acteurs, que ce soit pour financer de la recherche, de l'innovation ou du transfert technologique, sont nombreux et dispersés. Par ailleurs, **le niveau national ne semble plus être le niveau d'action pertinent pour ce type d'activités économiques :** les ministères manquent de moyens et de ressources humaines ; les moyens restants sont externalisés vers des agences indépendantes qui sont difficiles à coordonner (ANR, AII, OSEO-ANVAR, etc.) ; les règles communautaires interdisent de plus en plus le cofinancement européen et national rendant inutile voire dommageable l'intervention nationale. Dès lors, l'État doit reculer, mais en veillant à accompagner les acteurs. Cet accompagnement signifie prendre place au bon niveau d'intervention, coopération, expertise et, éventuellement, co-financement. Au niveau européen, il est essentiel d'organiser l'intervention, avec le SGAE, pour la conception des politiques, le suivi des actions, l'organisation de réseaux et de consortiums, et la réponse aux appels d'offres. Au niveau régional, il faut s'appuyer sur des structures compétentes assurant un rôle de portail telles que les pôles de compétitivité.

> Par les régions

Les **pôles de compétitivité** ont été créés pour concentrer sur une zone géographique les porteurs publics et privés d'innovation et améliorer les synergies, d'une part entre les différents éléments du dispositif de recherche et, d'autre part, entre ce dispositif et l'industrie du secteur. Divers pôles de compétitivité concernent les outils numériques, notamment **CapDigital** en région parisienne.

Accompagner les acteurs industriels, petits et gros, dans des démarches d'innovation n'est pas du ressort du ministère de la Culture et de la Communication. Par contre, il est important que le ministère participe à la vie de ces pôles afin que les problématiques culturelles et linguistiques y soient prises en charge sérieusement, par exemple en faisant ouvrir des lignes spécifiques sur les technologies de la langue dans les appels à propositions. D'un côté, le ministère pourrait **participer aux instances dirigeantes et aux commissions de choix des projets** en abondant les fonds et en envoyant des experts. Cela pourrait aller jusqu'à fusionner les instances d'évaluation d'un pôle comme CapDigital avec celle du RIAM afin d'épargner aux PME, au moins celles de la région, la réalisation d'un dossier supplémentaire. Cela permettrait aussi de communiquer sur les résultats de recherche et d'innovation atteints dans les secteurs culturels et linguistiques pour se faire connaître auprès des clients potentiels (les grands groupes privés notamment). D'un autre côté, le ministère pourrait avoir **un rôle de stimulation et de coordination des acteurs culturels participant à ces pôles.** Il est important que l'ensemble des grands établissements publics culturels y participent, notamment la BnF, le Louvre, la Cité des sciences et de l'industrie, la RMN, l'AFP, etc. (seuls l'INA, l'IRCAM et le Centre Pompidou sont actuellement actifs à CapDigital).

Parallèlement, devant la complexité des procédures et la nécessité de répondre en anglais, les acteurs du secteur, même les plus gros comme l'AFNOR, rencontrent de grandes difficultés pour **répondre aux appels d'offres communautaires.** Le ministère n'a pas les moyens d'aider les acteurs du secteur à monter des consortiums entre les établissements publics, les laboratoires du secteur et les entreprises. Avec les pôles de compétitivité, l'appui du SGAE et d'organismes tels que les Relais Culture Europe ou le réseau des Euro info centres, il serait fort utile d'imaginer **la création d'un bureau d'appui** sur le modèle du CNRS qui dispose « d'ingénieurs Europe », voire d'une personne morale susceptible de porter les projets de recherche et innovation des secteurs culturel et linguistique.

Par ailleurs, il existe un grand nombre de **guichets nationaux.** Les pôles de compétitivité remplissent **une fonction très utile de re-direction des acteurs vers ces guichets.** Rien n'interdit au ministère de flécher certaines aides en introduisant **une ligne spécifique « langue française » ou multilinguisme** grâce à l'abondement des fonds existants (par exemple le concours annuel d'entreprises innovantes de OSEO-ANVAR ; le dispositif commun OSEO / CNC sur l'audiovisuel et le multimédia ; le DICRéAM).

Enfin, mentionnons que les DRIRE (les services déconcentrés du ministère de l'Industrie)

gèrent des **aides collectives sectorielles pour l'industrie** (des « contrats de progrès » par exemple). Il pourrait être envisagé une action collective sur l'industrie de la langue qui déboucherait sur des actions comme de la formation, de la sensibilisation, de l'organisation de salons.

> Par l'Union européenne

Il apparaît que **le dossier « technologies de la langue » au niveau communautaire est très dispersé** entre la direction générale Société de l'information et médias (le volet technologies de l'information et de la communication du programme cadre recherche et développement, le programme cadre Compétitivité et croissance et la Bibliothèque numérique européenne), la direction générale Éducation et culture et le nouveau commissaire au multilinguisme. **Il est piloté uniquement par des unités recherche ou technologique sans réelle coordination avec les unités traitant des usages ou des impacts culturels.** Cette caractéristique se retrouve au niveau des correspondants nationaux (SGAE, ministère de l'Industrie principalement). Dès lors, les enjeux culturels et linguistiques qui motivent les interventions du ministère de la Culture et de la Communication sont extrêmement difficiles à appréhender. Assurant sa participation à la coordination des dossiers, le ministère pourrait, via le SGAE et le SGDN, sensibiliser les services du Premier ministre sur le caractère transversal du dossier de la langue, impliquant l'ensemble des directions générales de la Commission et l'ensemble des ministères français.

Le ministère pourrait trouver avantage à **organiser un GTN, groupe thématique national, sur les industries de la connaissance et/ou les industries culturelles.** Sur le modèle des GTN existants comme celui sur les TIC du ministère de l'Industrie ou ceux du ministère de la Recherche, son but serait d'animer le secteur afin d'informer de façon privilégiée les acteurs des échéances à venir, de faire remonter les demandes et d'aboutir à des demandes consensuelles à porter auprès des instances communautaires ainsi que d'harmoniser les positions françaises sur les grands dossiers bruxellois. Par ailleurs, le ministère pourrait utilement participer aux GTN industrie et recherche sur les TIC afin d'y porter les problématiques culturelles et linguistiques qui y sont trop souvent absentes.

Enfin, si les actions nationales ont naturellement leur utilité, des actions multinationales sont parfois plus appropriées, surtout dans la recherche, en particulier lorsqu'elles visent des évolutions de l'internet, qui n'ont pas d'impact strictement national. Le partage de l'effort entre les États-membres et la Commission semble très naturel, bien en phase avec le concept de subsidiarité et illustrant parfaitement la plus-value que peut apporter une coordination européenne sous l'égide de la Commission. Les États-membres assureraient en priorité la disponibilité de ressources (corpus, lexiques, dictionnaires…) en quantité et en qualité suffisantes, et la Commission, aurait pour mission première de coordonner l'effort, d'évaluer les performances des technologies et de veiller à définir des standards pour échanger les données. C'est pourquoi **les chercheurs français plaident de longue date pour la constitution d'un ERAnet+, d'un « article 169 », etc. Ces appels, bien que portés par le monde de la recherche, devraient être stratégiquement appuyés par le**

ministère de la Culture et de la Communication compte tenu des enjeux culturels du développement de ces technologies.

Assurer la promotion de l'utilisation des applications innovantes

Avoir une ambition économique et culturelle internationale implique d'être à la pointe à l'intérieur de nos frontières. Cela signifie bien évidemment être forts dans le développement des applications innovantes, mais également dans leur usage. **L'administration doit montrer l'exemple dans l'utilisation des outils logiciels linguistiques :** une politique dynamique d'**achats publics pré-compétitifs** des outils les plus innovants (si possible français plutôt que des outils commerciaux standardisés américains) à l'instar du moteur de recherche acheté pour le guichet unique patrimonial du ministère de la Culture et de la Communication ; une politique dynamique de **traduction des sites internet publics** (comme le prévoit la loi dite Toubon), des sites culturels en particulier, comme ceux des grands établissements publics en particulier ceux qui ont une vocation internationale comme l'INA, à l'instar du site du Louvre ; **la promotion des logiciels libres** respectant les caractéristiques culturelles et linguistiques (jeux de caractères, interfaces, documentation, aide en ligne) [DGLFLF / AFNOR, 2006] ; la veille et **la mise à disposition d'outils innovants type wikis**[11] (pour les contenus culturels et artistiques, les commissions de terminologie, etc.), **moteurs de recherche sémantiques** sur des contenus culturels.

Parallèlement, le groupe de travail sur la traduction dans l'administration coordonné par la DGLFLF préconise de développer une politique de traduction structurée au sein de l'État transposant les bons côtés de l'exemple canadien. Le bilinguisme constitutionnel du Canada implique une intense activité de terminologie, de normalisation et de traduction entre les deux langues constitutionnelles (l'anglais et le français), dont la mission incombe pour la plus grande part au **Bureau de la traduction.** Autour de cette matrice remarquablement organisée, une industrie langagière s'est constituée (formation aux langues, édition de méthodes, recherche et développement dans les technologies nouvelles…), qui représente une part importante de l'activité de services canadienne. Ainsi le Canada capte-t-il 10% de l'activité mondiale de traduction. La création d'une telle structure serait fort utile. Celle-ci serait exemplaire en matière d'achat, de développement et de commandes des outils nouveaux d'aide à la traduction, de participation aux activités de normalisation, de mise en réseau des ressources linguistiques permettant une consultation par le public, spécialisé ou non, de banques de données indispensables au dialogue multilingue. Le centre national du livre, la DGLFLF qui assure la gestion du fonds Pascal, les associations professionnelles de traducteurs, les établissements d'enseignement du métier de traducteur, déjà très actifs, trouveraient un interlocuteur étatique naturel dans cet organisme.

---

[11] Définition de Wikipedia : un wiki est un système de gestion de contenu de site internet qui rend les pages internet librement et également modifiables par tous les visiteurs autorisés. Les wikis sont utilisés pour faciliter l'écriture collaborative de documents avec un minimum de contrainte.

Pour ce qui concerne le secteur privé, **un dispositif équivalent** et donc exemplaire en matière d'utilisation des outils linguistiques informatiques, pourrait être **mis en œuvre par les chambres de commerce et d'industrie,** appuyé sur le réseau des chambres de commerce « franco-étrangères », afin de fournir des services de traduction aux PME-PMI. Il s'agirait bien évidemment d'aider l'offre privée de traduction à se structurer et d'en faire la promotion auprès des entreprises, et non de faire une concurrence déloyale aux traducteurs et aux entreprises de traduction.

Quelques grands projets qui font l'actualité du secteur

Au-delà des grandes recommandations de fond qui nécessitent de consacrer des ressources humaines et des financements qui parfois manquent, il semblerait judicieux de profiter de quelques dossiers d'actualité au cœur desquels se trouvent les technologies de la langue afin de revaloriser cette problématique et mettre en avant tous ses enjeux culturels et linguistiques.

> Européana, le projet de Bibliothèque numérique européenne (BNuE)

Le multilinguisme est au cœur de la problématique de la création d'une bibliothèque numérique européenne. L'accessibilité en ligne du matériel dans toutes les langues européennes constitue à la fois une force du projet et un enjeu de taille pour favoriser l'accès à la diversité et à la richesse du patrimoine culturel de l'Europe. Dans cette perspective, ce qui est déterminant n'est pas tant le volume numérisé que la capacité d'un utilisateur non spécialisé à retrouver les ressources disponibles. La conception et la combinaison de moteurs de recherche et d'outils de traduction capables de dépasser les barrières linguistiques, par une traduction des termes de l'interrogation et de la réponse, devient alors une nécessité.

Ce grand projet est pour le secteur des technologies de la langue une occasion de se développer en termes d'usages, de technologies et de recherche. Il est clair que seront achetés ou développés pour l'occasion des outils de recherche d'information interlingues, de traduction automatique ou assistée, de mise en correspondance de textes bilingues ou multilingues ou de correction orthographique permettant d'améliorer la reconnaissance de caractères dans les tâches de numérisation, etc. Ce chantier sera aussi une belle occasion d'encourager les standards qui permettent l'échange de données et leur archivage pérenne, facilitent les connexions avec les outils linguistiques existants et leurs développements, etc. ; de développer les ressources linguistiques (corpus, lexiques, bases terminologiques) en quantité suffisante pour pouvoir développer des technologies de qualité.

Pour le ministère de la Culture et de la Communication, dans ses différentes composantes, il est simple, légitime, utile et opportun d'investir sur ce projet tant du point de vue rhétorique, humain que financier. Même si aujourd'hui la réflexion semble quelque peu bloquée par des questions de principe comme le droit d'auteur ou les choix éditoriaux, ce « grand travaux » est l'un des seuls cadres européens de réflexion possibles

sur le développement des technologies de la langue, leurs applications, leurs usages et les impacts culturels et linguistiques.

Par ailleurs, il est essentiel que la priorité actuelle donnée à l'écrit, en particulier du côté français, n'entache pas **les développements dans le monde de l'audiovisuel et du multimédia.** En effet, l'INA est l'un des fleurons de notre secteur culturel public et il est essentiel que tant sa visibilité internationale - considérable - que son expertise soient mises au service de ce projet culturel, même s'il est originellement porté par le monde des bibliothèques.

> Le programme d'innovation industrielle Quaero

Quaero est un projet français avec une participation allemande. Déposé dans le cadre de la nouvelle agence pour l'innovation industrielle (AII), **le programme Quaero vise à l'organisation d'une filière industrielle portant sur le développement de nouveaux systèmes de gestion des contenus numériques multimédias pour des applications grand public ou professionnelles.** Il porte sur la conception et la réalisation de solutions nouvelles dans le domaine de la recherche d'informations numériques multimédias et multilingues en se concentrant sur les technologies du traitement automatique de la parole, du langage, de la musique, de l'image et de la vidéo. Le programme Quaero va contribuer au développement des technologies de la langue dans plusieurs directions : reconnaissance de la parole, du locuteur et de la langue, traduction, systèmes de question / réponse, etc. Il va aussi contribuer à la création de nouveaux outils permettant la numérisation des bibliothèques.

Le projet Quaero, en tant que programme d'innovation industrielle, repose avant tout sur le désir d'aboutir à des avancées technologiques. C'est pourquoi **il est important d'associer à ce projet des acteurs venant du monde culturel, portant une expertise sur les usages et sur les impacts linguistiques, culturels et symboliques.** Le ministère de la Culture et de la Communication a donc un rôle essentiel à tenir dans cette entreprise. *A fortiori* puisque sur l'internet, l'innovation est collective et incrémentale, et les utilisateurs de services répartis dans le monde entier y jouent un rôle essentiel.

Réciproquement, le ministère de la Culture et de la Communication a beaucoup à attendre de ce projet ambitieux qui pourrait, à moyen terme, prendre une part significative de la recherche d'informations sur le web et sur des réseaux propriétaires de contenus. **Au-delà de la participation de deux grands organismes culturels que sont l'INA et la BnF, le ministère pourrait s'associer directement au programme pour développer des applications supplémentaires** pour des projets comme le guichet unique patrimonial. Cette association pourrait prendre la forme d'un partenariat dans lequel le ministère apporterait son expertise et financerait le développement de corpus représentatifs des applications visées, et certains laboratoires et industriels du programme Quaero développeraient les systèmes correspondant à ces applications.

De façon plus générale, il est opportun de se poser la question d'une autre initiative dans le secteur des technologies de la langue, sur le modèle de Quaero : une agrégation de partenaires européens, une évaluation systématique des résultats, des financements publics à défaut de capitaux-risqueurs européens acceptant de jouer le rôle de financeur de projets d'envergure, etc. Les sujets les plus fréquemment avancés pour une autre initiative sont les deux autres grandes applications des technologies de la langue que sont la traduction d'une part et les interfaces hommes / machines d'autre part. Toutefois, pour le ministère de la Culture et de la Communication, cette question est tout à fait théorique dans l'état actuel de la situation.

Enfin, il est important de mentionner que malgré le fort effet mobilisateur du projet Quaero sur le secteur des technologies de la langue, les retombées positives attendues pour les PME choisies qui pourront profiter de façon privilégiée de certains résultats de la recherche publique et du « parapluie » des industriels intégrateurs, le projet ne profite pas à l'ensemble des acteurs. **Cette manne financière est loin de résoudre le problème du financement de la recherche sur les technologies de la langue.**

> Les guichets d'accès aux bases de données patrimoniales

Le ministère de la Culture et de la Communication s'est lancé dans différents projets exemplaires d'accès à un patrimoine documentaire numérisé grâce à des applications linguistiques innovantes. Citons en deux : le guichet unique d'accès aux bases de données patrimoniales et le portail multilingue de la documentation scientifique des musées ethnographiques exotiques.

Le comité interministériel pour la société de l'information (CISI) tenu en juillet 2006 sous l'égide du Premier ministre a décidé la création en 2007 d'un service permettant l'accès gratuit « en un clic » aux 16 millions d'œuvres et ressources numérisées dont 3,5 millions d'images par le ministère de la Culture et de la Communication. **En association avec ses établissements publics, le MCC a réuni 31 bases pilotes qui seront mises en ligne avec un moteur de recherche commun dès avril 2007** sous l'égide du haut fonctionnaire des systèmes d'information et du secrétariat général du ministère. En effet, depuis de longues années, le ministère produit des données d'une grande richesse, qu'il s'agisse de publications électroniques, d'inventaires ou bien sûr de patrimoine numérisé, mais dont l'accès est cloisonné et reste conçu pour des publics spécialisés ou avertis. Ainsi il existe plus de **240 bases de données** (Joconde, Mérimée, etc.), de sites internet (célébrations nationales, sites archéologiques, par exemple) ou encore 250 œuvres multimédias. **Il reste du chemin à parcourir.** Dans les deux prochaines années, sera accompli **un travail très important sur les normes de description des objets patrimoniaux et aussi sur l'élaboration d'un vocabulaire commun à l'ensemble de ces professions** pour la rédaction des méta-données, des systèmes de description, etc. Seule une rédaction homogène des bases permettra d'aboutir à des recherches sémantiques fines. **Il est important que la DGLFLF accompagne ce travail essentiel de terminologie** (trouver des définitions exactes et consensuelles) en s'appuyant sur les compétences professionnelles et scientifiques des

métiers patrimoniaux du ministère (archives, musées, inventaire, bibliothèques et, demain, le cinéma et l'audiovisuel). Par ailleurs, un partenariat avec les membres du consortium Quaero est à l'étude pour permettre des recherches sur des documents multimédias. Il s'agit d'un travail à accomplir rapidement compte tenu du rythme accéléré des plans de numérisation du ministère, des établissements culturels et des collectivités locales.

Reposant sur un tout autre choix technique et épistémologique (sans recours à une quelconque forme de normalisation des contenus), le musée du quai Branly, accompagné de la direction des musées de France et du CNRS, pourrait se lancer dans la réalisation d'**un portail multilingue de la documentation scientifique des musées ethnographiques exotiques.** Afin d'optimiser l'exploitation des fiches muséographiques qui ont été numérisées lors de la phase de chantier des collections du musée du quai Branly, la direction de la recherche du musée a souhaité mettre en application les résultats de la recherche sur la théorie des graphes développée pour traiter les grands corpus lexicographiques, aux catalogues de musées d'ethnographie exotique. L'occasion est double : avancer sur des recherches innovantes en sciences de la documentation et en ingénierie linguistique ; transcender les difficultés lexicographiques posées par la numérisation et la mutualisation des fiches muséographiques et, en particulier, le problème majeur de l'ethnonymie (noms propres) particulièrement sensible dans le secteur de l'ethnologie exotique. Par ailleurs, il a été envisagé de former un consortium européen autour de ces travaux. Cela permettrait d'avoir accès à d'autres bases de données de même type et d'autres expertises, de mettre le projet à l'épreuve du multilinguisme et, éventuellement, d'obtenir des fonds communautaires au titre de l'un des programmes cadres de la Commission.

Sans vouloir multiplier les exemples, sans insister non plus sur les applications innovantes développées depuis bien longtemps dans des établissements culturels tels que la BnF ou l'INA, rappelons seulement l'importance de ces projets qui, bien que multiples, concourent tous, d'une part, à un meilleur accès aux données publiques culturelles et, d'autre part, à la mise en pratique par le ministère lui-même, de son devoir de stimulation de la recherche et de l'innovation et de l'achat pré-compétitif d'applications linguistiques innovantes.

Engager la bataille du multilinguisme à l'occasion de la présidence française

Face à la menace d'un monolinguisme institutionnalisé où certains croient voir un idéal et d'autres une simple commodité, la polyphonie des langues de notre continent, partagée par plus de 450 millions d'Européens, est un héritage à promouvoir, à structurer et à transmettre. Pour l'Europe, la langue porte un double enjeu fondamental : à la fois veiller à préserver les cultures des différents pays membres, en permettant aux citoyens de s'exprimer dans leur langue maternelle (identité / diversité), mais également faciliter la communication entre les citoyens de ces différents pays, parlant différentes langues. Aujourd'hui, au niveau communautaire, la question linguistique n'est ni secondaire ni simplement pratique, elle est devenue fondamentale.

Pour répondre à ce défi, tous les instruments de dialogue et d'échange entre les langues

dont l'Europe s'est dotée au fil des années doivent être profondément valorisés, la traduction en particulier. Cependant, la traduction est aujourd'hui un domaine largement négligé par l'Union, qui n'aborde en général cette question que sous l'angle pratique de l'interprétation des réunions ou de la traduction des documents pour ses besoins propres de fonctionnement auquel elle consacre d'ailleurs un budget considérable. La diversité linguistique étant vécue négativement, comme un obstacle à la communication, l'Union européenne éprouve quelque peine à la valoriser dans ses politiques. **La valorisation des outils de traduction automatique est un moyen de présenter la traduction comme un gain et non comme une charge.**

La présidence française de l'Union européenne au second semestre 2008 pourrait être l'occasion pour les autorités françaises, la Fédération européenne des institutions linguistiques nationales et la Francophonie de proposer une grande réflexion sur les enjeux du multilinguisme au travers des outils de traduction automatique. Le nouveau commissaire au multilinguisme pourrait être incité à prendre le pilotage d'une réflexion sur la langue et les nouvelles technologies, en insistant sur la dimension linguistique du dialogue inter-culturel, afin de dégager une importante valeur ajoutée par rapport à tout le travail que fait déjà la direction générale Société de l'information dans le domaine de la recherche et du développement. De grands projets tels que la BNuE sont des écots essentiels à apporter dans l'escarcelle du dossier « multilinguisme ».

Les principales mesures préconisées

Créer la fonction. S'approprier le dossier des industries de la langue signifie pour le ministère la création d'un poste et le recrutement d'un chargé de mission de haut niveau. À défaut d'une délégation aux industries culturelles encore inexistante assurant une veille sur les industries de la connaissance (et donc aussi de la langue) et participant activement à la coordination intra et interministériel, la DGLFLF comme la DDAI, sont susceptibles d'assurer ce rôle.

Militer pour un Technolangue 2. La DGLFLF et la MRT doivent ensemble instruire un dossier Technologies de la langue pour le porter à l'ANR comme l'une des priorités du MCC qui contribuera à ce programme par des financements et par son expertise.

S'appuyer sur une instance de coordination et d'impulsion. Soit en créant un nouvel organisme, soit en assignant cette fonction à un organismes existant, la création ou l'assignation d'une instance sera en soi un acte fort de légitimation du dossier par les pouvoirs publics.

Mettre à disposition des ressources linguistiques. Qu'il s'agisse de corpus ou de ressources plus complexes comme des dictionnaires, etc. cette activité incontournable ressortit à la responsabilité de l'État. Cette action doit être intégrée à la politique générale du MCC relative à la mise à disposition des données publiques culturelles essentielles.

Appuyer la présence française dans les chantiers internationaux de normalisation. Cela passe par un appui circonstancié à l'envoi d'experts français et francophones dans les instances internationales de normalisation, la promotion et l'application des normes existantes et le maintien du français comme langue officielle pour les normes ISO.

Accompagner la prise en charge de l'aide à la recherche et à l'innovation par les régions et par l'Union européenne. Le MCC doit s'impliquer, par ses financements, son expertise, son influence sur les établissements publics culturels, etc. dans des instances telles que les pôles de compétitivité, notamment CapDigital en région parisienne, et les programmes cadres de la Commission européenne.

Assurer un rôle de promotion de l'utilisation des applications innovantes. L'administration doit montrer l'exemple dans l'utilisation des outils logiciels linguistiques en ayant une politique dynamique d'achats publics pré-compétitifs et d'usage des outils les plus innovants. La multiplication des **projets de guichets d'accès aux bases de données patrimoniales** va dans ce sens.

Participer aux grands projets qui font l'actualité du secteur tels que **Européana,** le projet de Bibliothèque numérique européenne (BnuE) et le programme d'innovation industrielle Quaero.

Proposer le multilinguisme comme l'une des priorités de la France **pour la présidence de l'Union européenne** au second semestre 2008. Il est essentiel de faire passer le message que la diversité linguistique est, pour l'Europe, non pas une charge, mais une richesse sur laquelle s'appuyer culturellement et économiquement.

État des lieux

Les outils

Il n'existe que peu de définitions et aucune définition consensuelle de ce que l'on pourrait désigner de façon minimaliste comme « l'ensemble des applications informatiques utilisant des ressources linguistiques » ou bien, simplement, la manipulation à l'aide des outils informatiques du texte écrit et de la parole. Nous chercherons donc plutôt ici à offrir une description du secteur qu'une véritable définition.

Les points de rencontre entre les langages informatique et humain

Des terminologues zélés

La terminologie est floue et renvoie indifféremment à des objets, des processus, une discipline scientifique ou un secteur d'activité (fédération d'acteurs). Ce secteur d'activités, au carrefour de l'informatique et de la linguistique, est **désigné sous les noms les plus divers,** souvent en anglais, de « technologies de la langue » ou « technologies du langage » ou *« human language technologies »* (HLT), « industrie[12] de la langue » (IL), « industries des langues », « traitement automatique des langues » (TAL), « traitement automatique du langage naturel » (TALN), « traitement informatisé des langues » (TIL), « linguistique informatique » ou « informatique linguistique », « linguistique computationnelle », « ingénierie linguistique », « nouvelles technologies d'analyse de l'information » (NTAI) ou, très simplement, « outils linguistiques ».

La terminologie a évolué dans le temps. Dans les années 1990, le terme « industrie de la langue », démodé, a été remplacé par celui « d'ingénierie linguistique » faisant référence au « génie linguistique », cousin des génies civil, électrique, etc. Depuis la fin des années 1990, le mariage de la linguistique et de l'analyse statistique, le mariage de la recherche sur l'écrit et de la recherche sur l'oral ont donné naissance aux termes de technologies de la langue (TL) (*human language technologies* en anglais) et de traitement automatique des langues (TAL). Néanmoins, le TAL a dans certains contextes un sens plus étroit faisant référence uniquement aux approches sémantiques du traitement de l'écrit.

Le **terme « multilingue »** ou **« plurilingue »** ou « interlingue » est utilisé dans deux sens principaux. Dans un sens, il renvoie à des ressources ou **des outils monolingues, mais développés pour plusieurs langues.** Par exemple, un logiciel de résumé est capable de résumer en espagnol un texte en espagnol et en italien un texte en italien. L'enjeu est alors

---

[12] En tout état de cause, le terme industrie doit être aussi entendu dans son sens non pécuniaire comme l'ensemble des opérations concourant à la production des richesses par la transformation des matières premières, ici linguistiques.

que la conception informatique de l'outil soit la plus « ouverte » possible c'est-à-dire qu'elle permette d'utiliser l'outil avec un nombre toujours plus grand de langues différentes (ne pas oublier que certaines langues ont des accents ou ne s'écrivent pas en caractères latins par exemple). Dans l'autre sens, il renvoie à **des outils qui croisent les langues.** Un logiciel de résumé peut résumer en italien un texte en espagnol. L'enjeu est alors de se focaliser sur la combinaison des langues entre elles (506 paires – 23 x 22 - à la Commission européenne, par exemple). Dans l'accès aux contenus, plurilinguisme a aussi deux sens complémentaires : la traduction de la requête *a priori* qui permet d'interroger l'ensemble des éléments en plein texte ou des descripteurs, ou bien la traduction des éléments apportés en réponse à une interrogation monolingue.

L'inventaire

Grâce à l'organisation de la normalisation, on peut tenter d'approcher le paysage des aspects linguistiques de l'informatique. À l'ISO, le comité technique (TC) 37 s'intitule « Terminologie et autres ressources langagières et ressources de contenu ». Il s'agit du cœur de la « galaxie TAL ». C'est là qu'on retrouve les dictionnaires, les applications de traitement de la langue comme l'indexation ou le résumé, la traduction, les moteurs de recherche. Néanmoins, d'autres comités techniques et de nombreuses instances de normalisation discutent aussi d'autres points de rencontre essentiels entre l'informatique et la langue.

Il s'agit principalement :

> des activités de « translitération » et d'indexation des informations écrites concernent traditionnellement les métiers de la **documentation,** des bibliothèques et de l'archivage (TC 46 de l'ISO) ;
> du traitement des **contenus multimédias** (sous-comité SC 29 de l'ISO et par le W3C) ;
> des **corpus oraux** (Forum Voice XML du W3C, etc.) ;
> de **l'internationalisation des logiciels** (SC 22 de l'ISO) ;
> du **codage des caractères** et ses applications comme le DNS (principalement discutés par le consortium Unicode) ;
> des **interfaces utilisateurs** de tous les appareils informatiques (TC 1 / SC 35 de l'ISO) ;
> du web **sémantique** et **les schémas de portabilité comme le XML** (W3C) ;
> de la **terminologie et la lexicographie** (TC 37 de l'ISO).

Pour aller un peu plus avant dans **l'inventaire** du secteur, envisageons le détail du seul comité technique 37 de l'ISO « normalisation des principes, méthodes et applications relatives à la terminologie et aux autres ressources langagières et ressources de contenu dans les contextes de la communication multilingue et de la diversité culturelle ». Il **laisse voir la complexité et la variété des domaines à traiter :**

> le sous-comité 1 « Principes et méthodes » comprend : Harmonisation de la **terminologie,** Principes, Méthodes et vocabulaire, Socio-terminologie, Modélisation des

concepts dans le travail terminologique ;

- le sous-comité 2 « Méthodes de travail terminographiques et lexicographiques » comprend : Codage de noms de langues, Terminographie et lexicographie (fabrication des **dictionnaires**), Identification des sources pour les **ressources linguistiques,** Exigences et modèles de certification pour la gestion de la diversité culturelle, Services de **traduction** et **d'interprétation ;**
- le sous-comité 3 « Systèmes de gestion de la terminologie, de la connaissance et du contenu » comprend : Éléments de données, Vocabulaire, Échanges de données, **Gestion des bases de données ;**
- le sous-comité 4 « Gestion des ressources linguistiques » comprend : Mécanismes et descripteurs de base pour les **ressources linguistiques,** Schémas de représentation, Représentation de **données multilingues,** Ressources lexicales, Gestion des flux de données linguistiques.

Des technologies, entre connaissances et applications

Il est possible de **représenter le secteur par une chaîne** partant des corpus constitués (écrits, oraux, multimédias) allant vers des services génériques répondant à des besoins et/ou créant des usages (traduction, pédagogie, veille informationnelle, gestion documentaire) et passant par la constitution de briques linguistiques et de briques informatiques plus ou moins élaborées et plus ou moins imbriquées entre elles. En termes d'activité, la chaîne va de la recherche fondamentale (plus de connaissances) au développement de nouvelles applications (plus de services) en passant par l'invention ou l'amélioration d'outils et de technologies. En termes d'acteurs, la chaîne part des chercheurs, passe par des développeurs et se termine chez les intégrateurs. Et réciproquement !

	Connaissances	**Technologies**[13]	**Applications**
Quoi ?	Ressources linguistiques Résultats de la recherche en sciences du langage Résultats de la recherche en informatique, en statistique, en mathématique, en traitement du signal Résultats de la recherche en ergonomie cognitive et en psychologie	Modules informatiques Modules linguistiques Ressources linguistiques	Logiciels Progiciels Services
Comment ?	Recherche fondamentale et appliquée sur . La langue = écrit + oral . L'image . Le geste . La cognition . Le signal	Développement et évaluation de technologies	Intégration
Qui ?	Les laboratoires de recherche publics et privés	Les laboratoires de recherche publics et privés Des consortiums Des agences de mutualisation (ELRA en Europe, CNRS en France, LDC aux USA)	Les industriels petits et grands

[13] Le terme de « technologies » est entendu ici dans le sens de science des techniques.

Les activités de recherche

D'un point de vue scientifique, on peut considérer deux caractéristiques de ce secteur : les étapes de la recherche sont nombreuses ; les extrants sont très dépendants des intrants. D'un point de vue économique, il est nécessaire d'avoir en tête que la chaîne de valeurs est très découpée. À l'exception des maisons d'édition, l'activité principale des fournisseurs de technologies n'est pas la production, la collecte ou la validation des ressources linguistiques. En pratique, la plupart des fournisseurs développent en interne ou acquièrent les ressources, les lexiques ou les dictionnaires qui sont « appelés » par les applications informatiques, pour leurs propres besoins. Le marché des ressources linguistiques (pour une seconde exploitation) est donc un nouveau « marché », commercial ou non, mais en tout cas essentiel.

Afin de simplifier, on peut scinder la description de la recherche en plusieurs grandes catégories :

> **la recherche fondamentale** dans les disciplines généralement sollicitées pour développer ou intégrer des technologies de la langue (linguistique, mathématique, informatique, statistique, traitement du signal, sciences de la cognition et même géographie, économie, etc.) et évidemment en psychologie et sociologie afin de connaître l'état des usages ;

> la recherche appliquée aux **technologies ;**
> **les activités d'appui à la recherche** (la normalisation, la constitution et la mutualisation de ressources linguistiques, l'évaluation) ;
> la recherche sur **les applications** (la traduction, le traitement de l'information et les interfaces principalement).

Il serait nécessaire de disposer d'une cartographie des activités et des acteurs, de l'enchaînement des usages, des points de blocage, des activités déjà balisées par les scientifiques et les industriels et des activités encore « vierges », des activités pour lesquelles les Français (ou les Européens ou les Francophones) sont précurseurs... On se contentera ici de quelques éléments de contexte.

D'un point de vue général, il semblerait y avoir consensus sur l'idée que, depuis 50 ans, la recherche sur les technologies de la langue a plutôt évolué par capitalisation et **qu'il est illusoire d'envisager de rentrer dans une période de rupture scientifique.** Vraisemblablement, il continuera d'y avoir une multitude d'inventions, provenant principalement du secteur public, plutôt que de véritables innovations. Ce constat ne facilite pas la prise en charge politique du financement de la recherche.

À côté de la recherche appliquée, la nécessité de développer les activités d'appui à la recherche revient dans ce secteur comme un leitmotiv.

La constitution de ressources informatiques

Les modules informatiques dont il s'agit ici reposent donc tous sur des paramètres linguistiques. Il s'agit là de leur point commun. Par exemple, les lexiques servent à fabriquer les outils informatiques reposant sur le traitement de chaînes de caractères (correction orthographique, moteur de recherche comme Google par exemple). Les dictionnaires et les ontologies, au sens de réseaux sémantiques, sont utilisés dans le traitement automatique de la langue (TAL) à proprement parler (moteurs de recherche linguistiques, résumé automatique, veille informationnelle, traduction automatique). Ces modules informatiques traitent l'information sous forme écrite, orale ou multimédia, monolingue ou multilingue, etc. Ils sont identiques quelles que soient la ou les langues qui les alimentent.

La constitution de ressources linguistiques

« A noir, E blanc, I rouge, U vert, O bleu : voyelles,
Je dirai quelque jour vos naissances latentes... »
Arthur Rimbaud, 1883

Les ressources linguistiques sont tous les types de données relatives à la langue, orale ou écrite, accessibles dans un format numérique, et utilisées pour le développement des applications informatiques et la recherche. Ces ressources sont propres à chaque langue ou à chaque paire de langues et fortement dépendantes du domaine d'application. Il s'agit par exemple de corpus écrits ou oraux annotés, de lexiques, de lexiques avec prononciation, de dictionnaires, de thésaurus, d'ontologies, etc.

Les développeurs de technologies de la langue ont des besoins importants et incessants de ressources linguistiques pour alimenter et évaluer ces technologies. Il est difficile de convaincre des chercheurs de travailler sur la production de ces ressources de base, les corpus en particulier, car c'est un travail compliqué qui ne peut pas être accompli par des juniors non encadrés, c'est un travail ingrat et difficilement valorisable par un scientifique (via les publications). Par ailleurs, **il est important que les ressources existantes soient mises à la disposition de l'ensemble de la communauté scientifique et industrielle.** Cela nécessite des développements en **formats libres** (non propriétaires), des **instances de diffusion** (comme le LDC américain ou l'agence ELRA européenne), **des instances de concertation par langue** afin de bien choisir les ressources en fonction des besoins applicatifs et des besoins pour l'évaluation, **une politique de prix,** voire de gratuité, prise en charge politiquement.

On peut distinguer les deux sortes de ressources linguistiques que sont les corpus et les « briques » linguistiques (lexiques et dictionnaires, les bases de données terminologiques ou sémantiques, etc.).

Grâce au traitement (lexicologique [lexicographique et terminologique], phonologique et syntaxique) des corpus, on obtient des grands types « d'objets », de **briques linguistiques,**

chacun pouvant être monolingue ou multilingue. Leur qualité repose sur la richesse et la qualité des corpus utilisés. Ces outils sont aussi obtenus par numérisation d'objets lexicographiques et terminologiques existants.

Les principaux « objets » sont :

> les **lexiques** (ou index ou listes de chaînes de caractères) ;
> les **dictionnaires** (sémantiques) monolingues ou multilingues. Les dictionnaires, par exemple ceux qui sont intégrés aux logiciels de traduction, ne sont pas seulement une liste de mots ou d'expressions avec leur traduction. Chaque mot ou chaque expression doit être défini avec des **informations linguistiques (morphologie, sémantique, syntaxe, style...)** dans la langue source puis dans la langue cible. Ces informations sont ensuite gérées par le moteur de traduction. Plus les dictionnaires sont riches, plus la traduction obtenue est précise ;
> les **thésaurus** sont des listes hiérarchisées de termes représentatifs d'un corpus ;
> les **réseaux sémantiques** ou les **réseaux conceptuels** fabriqués par appariement d'un dictionnaire général et de thésaurus « métiers » (spécifiques à un secteur d'activité) dans lequel chaque mot correspond à une idée et est relié aux autres mots. Ces réseaux sémantiques sont communément appelés **« dictionnaires ».**

Certains chercheurs constituent aussi des **ontologies.** Il s'agit d'une classification intellectuelle d'un domaine, voire du monde entier pour les plus ambitieuses, comme celle de l'université américaine de Princeton (www.wordnet.com). Chaque ontologie est propre à une langue, voire à la culture qui la produit. Depuis quelques années, il existe aussi des techniques de fabrication des ontologies par les utilisateurs eux-mêmes, alors appelées « folksonomies », grâce à un système de liens et d'étiquettes (balises, tags), le « *tagging* ». Ces possibilités propres à ce qu'il est désormais convenu d'appeler le web sémantique viennent inverser totalement la problématique de la construction ontologique.

Les activités d'appui à la recherche

> La mise à disposition de corpus

Les corpus permettent, en amont, de fabriquer des modules linguistiques qui leurs sont propres (par exemple un lexique spécialisé) à partir d'un ensemble de textes techniques et, en aval, à évaluer les applications (et bien sûr à conduire des recherche dans le domaine linguistique). Il existe 3 grands types de corpus monolingues ou multilingues :

- les corpus oraux
- les corpus écrits
- les corpus multimédias

La mise à disposition de corpus écrits et oraux est très importante pour la communauté scientifique comme pour les industriels. Chaque corpus élaboré par un linguiste peut servir aux autres linguistes même dans les autres spécialités de la discipline et peut servir à d'autres disciplines scientifiques (les historiens, les sociologues, les politistes, la

recherche médicale sur les maladies de la communication comme l'aphasie par exemple, etc.). Chaque corpus peut aussi servir aux industriels qui développent des technologies. Certains corpus servent plus particulièrement à réaliser des campagnes d'évaluation des technologies.

La mise à disposition actuelle des corpus oraux, parce que d'une grande actualité, montre certains enjeux de la recherche : l'accès au plus grand nombre de larges corpus, les différents usages scientifiques qui en sont faits, le développement de logiciels (ingénierie linguistique) les utilisant, le développement d'applications et de services finaux aux utilisateurs. Malheureusement, **la mise à disposition de corpus en français, oraux en particulier, est en retard.** Sur l'internet, on trouve déjà des corpus oraux en anglais, en espagnol, en portugais, etc. beaucoup plus nombreux. Néanmoins, la France rattrape doucement son retard et tente de profiter de ce rattrapage pour devenir exemplaire dans les modes de mise à disposition. Un groupe de travail s'est mis en place sous l'égide de la DGLFLF et s'attelle à la formalisation de différents procédés (balisage, indexation, transcription, etc.) qui sont des débuts de normalisation pour les corpus oraux. Il est essentiel que les spécificités de la langue française, et à travers elle de l'ensemble des langues non dominantes, puissent être prises en compte (les sons, la morphologie des mots et des phrases, etc.). Il manque encore des corpus en particulier sur des textes mal rédigés (comme des courriels - ou des minimessages *(sms)* contenant des fautes syntaxiques et orthographiques) ou mal prononcés.

> L'évaluation

En 1985, les Américains ont mis en place un système d'**évaluation** des résultats de la recherche technologique selon un protocole précis. Il permet de comparer les résultats des méthodes et des équipes, de mesurer les progrès de la technologie et donc de rentabiliser des investissements. En 1992, les campagnes annuelles d'évaluation se sont ouvertes aux laboratoires étrangers. Certains laboratoires européens publics et privés ont répondu et ont obtenu de bons résultats. Les tâches évaluées sont de plus en plus complexes. En 2006 par exemple, il s'agissait de transcrire, résumer et indexer un journal télévisé en plusieurs langues (vocabulaire illimité, plurilingue, pluri-locuteur).

Certaines campagnes d'évaluation ont été lancées par l'AUPELF (AUF) dans l'espace francophone et en France dans le cadre du programme Technolangue. Ces campagnes sont onéreuses (environ 250 000 euros par campagne en 2007), mais permettent de rentabiliser des investissements bien plus importants (plusieurs millions d'euros). Pour le moment, **il n'existe d'équivalent ni français ni européen du système américain d'évaluation** qui résulte d'une forte volonté publique (DARPA / NIST[14]). Il reste donc à inventer un tel système permettant aux pouvoirs publics de mesurer la qualité des résultats grâce à des objec-

[14] L'agence américaine en charge des projets de recherche avancée dans le domaine de la défense (équivalent de la DGA) et l'organisme américain de mesures et d'essais (équivalent du LNE).

tifs et des indicateurs définis *a priori*[15]. La délégation générale pour l'armement du ministère français de la Défense affiche cette activité comme une priorité pour les prochaines années.

> La normalisation

Au contraire de la règlementation, la normalisation est volontaire. Une norme représente un niveau de consensus entre experts d'un domaine spécifique et n'est proposée qu'à titre d'option possible. Une norme naît généralement d'une habitude initiée par un pionnier (comme les référentiels de bibliothèques par la France par exemple) puis est adoptée par propagation puis interceptée par les instances de normalisation pour discussion et formalisation. **La normalisation est essentielle, car elle est un lien naturel et important entre l'industrie** (les utilisateurs) **et les chercheurs,** et elle permet de réaliser une veille technologique et une veille normative très efficaces et diffusables vers les industriels. En outre, il est important de noter que **la normalisation est essentielle dans le domaine du logiciel libre** puisqu'il est important de pouvoir mettre en relation de travail ou substituer les uns aux autres les différents logiciels. Or, le secteur européen des technologies de la langue est profondément lié au logiciel libre : il est principalement porté par le secteur public de la recherche ; l'Europe ne dispose pas de grands groupes fabricants de systèmes informatiques propriétaires (comme Microsoft) ; les technologies sont très « diffusantes ».

On distingue généralement **les normes *de jure* des normes *de facto.*** Une norme *de facto* désigne une norme qui a été élaborée par une organisation autre qu'un organisme formel de normalisation. Une norme *de facto* peut provenir de plusieurs sources d'activités, variées par le type d'influence technologique et par la nature diverse de leurs intérêts. Ainsi, certaines normes sont établies *de facto* par un consortium de fabricants d'ordinateurs, un consortium de vendeurs, un regroupement de concepteurs de logiciels, une grande entreprise informatique, un groupe d'utilisateurs, une association professionnelle, un organisme responsable de politique d'acquisition de technologies, un organisme de financement de la R&D, un groupe d'évaluation de technologies. Le dernier inventaire [AFNOR] avait recensé près de 800 organismes différents sur ce secteur. Une norme *de jure* (ou formelle) désigne une norme qui a été établie par un organisme légalement (ou formellement) constitué et mandaté pour élaborer et développer des normes. La normalisation se fait alors à deux niveaux : au niveau national, à l'AFNOR ; au niveau international, à l'ISO.

Globalement, **trois grands secteurs d'activités de normalisation** *(de jure* et *de facto)* nous intéressent davantage : **le traitement informatique de la langue française** (TILF), **le traitement informatique de langues naturelles** (TILN), **les technologies du traitement de l'information** (TI) dont le nouveau courant est identifié par le terme « **internationalisa-**

---

[15] Il faut aussi penser cette évaluation comme un complément pragmatique, voire un support, au système des publications comme échelle de reconnaissance de la valeur scientifique.

tion »[16]**.** À l'ISO, deux grands comités techniques (TC) sont concernés par les outils linguistiques : le comité technique 46, qui concerne les métiers de la documentation, des bibliothèques et de l'archivage (ce comité technique est actuellement présidé par une représentante de la BnF et son secrétariat assuré par l'AFNOR) ; le comité technique 37 qui concerne la terminologie et la lexicographie (voir le paragraphe « L'inventaire » de ce chapitre pour plus de détails).

Depuis plusieurs années, **le français n'est plus une langue de travail de l'ISO,** même s'il reste l'une des trois langues officielles (avec l'anglais et le russe ; les Chinois plaident d'ailleurs pour que le mandarin remplace le français). L'ISO a externalisé vers l'AFNOR la traduction des normes qui les traduit une fois validées en anglais. Avec la traduction, l'AFNOR remplit une mission de service public, dirigée en particulier vers les PME, les associations professionnelles, les associations de consommateurs, les collectivités locales, les ingénieurs francophones du sud... qui ne peuvent pas toujours comprendre les normes rédigées en anglais.

La normalisation française présente des difficultés conjoncturelles à cause du désengagement des grands groupes et du tarissement relatif de l'innovation qui est le substrat normal de la normalisation. Dans le secteur informatique, la normalisation *de jure* est accusée d'être trop lente (en moyenne 30 mois pour faire une norme) et les normes s'établissent plutôt *de facto* dans une myriade de différents organismes (plusieurs centaines). La représentation de chaque pays dépend de sa capacité à envoyer des experts et du niveau de compétence de ces experts. Dans le secteur informatique, la plupart des membres actifs viennent du secteur commercial nord-américain. Même si ceci n'affecte pas la validité technique des décisions prises, cela laisse néanmoins penser que les contraintes et les besoins des autres parties du monde ne sont qu'insuffisamment prises en compte[17]. Comme l'évaluation, la normalisation prend beaucoup de temps et doit être suivie par des professionnels accomplis ; elle ne permet pas non plus de valorisation par la publication. Probablement par manque de culture du lobbying, il n'y a pas de Français « professionnels » sectoriels de la normalisation dans les instances internationales. Les experts rencontrent trop fréquemment des problèmes de reconnaissance de cette activité par leur employeur, de décharges de travail et de frais de mission. L'AFNOR n'a pas de budget pour envoyer des missionnaires à l'ISO et de moins en moins d'argent pour recruter ses propres spécialistes.

---

[16] L'ISO a défini ce terme comme étant un processus de production d'une application qui aurait la capacité d'être utilisée dans plusieurs environnements nationaux (ou culturels) de manière à ce qu'elle puisse produire une sortie et lire une entrée dans un format approprié pour chacun de ces environnements. L'internationalisation ne se limite pas uniquement aux propriétés multilingues d'une application, elle implique aussi toute différence entre les cultures, les coutumes et les habitudes et veille au respect des exigences socioculturelles des utilisateurs des technologies (le respect des caractéristiques des différentes langues nationales, le principe d'un environnement de travail dans sa langue maternelle, le droit de poursuivre et de faire valoir sa propre entité culturelle).

[17] Quand l'internet a commencé à se développer en France, les premiers utilisateurs du courrier électronique se sont rapidement émus de ne pouvoir utiliser l'intégralité des caractères typographiques du français et de recevoir des codes incompréhensibles à la place des « é », « è » et autres « ç ». Les chercheurs qui ont créé le protocole de courrier « SMTP » aux États-Unis, il y a une vingtaine d'années, n'avaient besoin que des caractères de l'anglais.

Le problème des accents dans le courrier électronique, encore imparfaitement résolu, a révélé aux non-spécialistes de ces questions **l'importance et les enjeux des normes et des standards pour les langues et les cultures,** non seulement pour le courrier électronique, mais aussi pour la structuration des documents, la définition des claviers et des interfaces, l'internationalisation et la localisation des logiciels. Des normes techniques peuvent entraîner des conséquences majeures par exemple sur la sécurité, la confidentialité, le multilinguisme ou la propriété intellectuelle, qui relèvent de questions sociétales. En matière de multilinguisme sur internet, la maturité et la stabilité des protocoles ne sont pas aujourd'hui suffisantes pour qu'ils soient reconnus comme protocoles internationaux incontestables. À l'occasion de la présentation du projet de Bibliothèque numérique européenne par la France aux instances communautaires, des pistes de travail ont été annoncées dans le domaine de la normalisation : des formats de méta-langage et de méta-données, d'archivage, de services web (moteur de recherche, interface de consultation), de l'interopérabilité d'échanges de données, etc. En effet, il devient urgent de disposer d'une norme de codage des contenus multimédias et multilingues afin de développer des outils communs aux différents pays concernés par le projet de bibliothèque.

La fabrication des technologies, entre recherche et intégration

Les grandes familles d'outils

Il existe des grandes « familles » d'outils. Chaque outil peut être mono ou multilingue. Ils correspondent aux quatre grands niveaux de l'analyse linguistique : le niveau morphologique (la reconnaissance du mot), le niveau lexical (la lemmatisation ou réduction du mot à sa forme canonique), le niveau syntaxique (avec utilisation de la grammaire) et le niveau sémantique (avec reconnaissance des concepts).

Les outils **d'analyse de la langue** et de **génération de la langue :**

- pour certaines langues, les outils de **segmentation**[18] reconnaissent les segments (les phrases ou les mots) qui composent le texte ;
- après l'étape de segmentation, les outils **d'étiquetage** (*tagger*) reconnaissent les formes particulières telles que le nombre, le genre, la personne, des mots, et attribuent des étiquettes morpho-syntaxiques aux mots ;
- les outils de **lemmatisation** reconnaissent la forme canonique d'un mot et ramènent chaque terme à une forme unique (l'infinitif des verbes, le masculin singulier, etc.). L'analyse lexicale consiste à ramener les mots à une forme de base, le radical, la racine ou le lemme, et à reconnaître toutes les variations liées à cette forme ;
- les outils de **détection des entités nommées ;**
- les outils des **détection des thèmes ;**
- le **correcteur orthographique** est généralement intégré à un logiciel de traitement de texte ;

[18] La segmentation obéit à des règles complexes basées en principe sur les espaces et la ponctuation.

- les outils **d'analyse syntaxique** reconnaissent les relations de différentes natures que les mots entretiennent entre eux, comme celles qui existent entre un verbe, son sujet et ses compléments, un nom avec son adjectif ou son déterminant, etc. (cf. le **correcteur syntaxique** intégré au logiciel de traitement de texte) ;
- les outils de **l'analyse sémantique** mettent en relation un texte avec une structure (une base de connaissances) représentant le sens des mots. Cette base de connaissances peut inclure les synonymes, les équivalents avec une autre langue et les termes génériques ou spécifiques (il s'agit alors d'une ontologie). Cette analyse est plus ou moins **pragmatique** c'est-à-dire qu'elle tient plus ou moins compte du sens différent d'un terme pour un utilisateur particulier (ou une catégorie d'utilisateurs) ou pour une situation particulière (le contexte dans lequel on se trouve) ;
- le générateur de **résumé ;**
- le générateur de textes à partir d'une base de données ou à partir d'un résumé.

Ce niveau d'analyse est utile à l'ensemble des champs applicatifs : dans le champ de la recherche d'information par exemple où, pour l'essentiel, il continue l'usage traditionnel des thésaurus ou des listes d'autorité et dans celui de la traduction automatique où il permet d'établir le lien entre plusieurs langues.

Les outils de **connaissance et de renouvellement de la langue :**
- la lexicographie formalisée et les **dictionnaires** électroniques ;
- les outils **d'alignement** de textes (pour construire automatiquement des dictionnaires multilingues et des mémoires de traduction) ;
- les **extracteurs de terminologie.**

Certains outils apportent un service pour le **traitement de l'information non structurée.**

Les outils de la **gestion** ou de l'analyse **de documents** (rechercher, filtrer et extraire dans une grande masse de documents) :
- les outils d'**indexation automatique de l'écrit.** Selon la définition classique, c'est-à-dire documentaire, l'indexation est la représentation par les éléments d'un langage documentaire des notions résultant de l'analyse d'un document ou d'une question en vue d'en faciliter la recherche. Selon l'approche linguistique de l'indexation qui renvoie à l'essor actuel des outils de traitement automatique du langage naturel (TALN), il s'agit de la représentation du document lui-même ;
- les outils d'indexation automatique **d'images fixes ou animées.** Cette indexation repose sur la présence ou l'absence d'un objet identifié par ailleurs. Si l'on compare avec le TAL, il s'agit du même processus que l'indexation d'une chaîne de caractères. Il ne s'agit pas d'une indexation « sémantique » ;
- les outils statistiques comme la recherche de co-occurrences (linguistique computationnelle).

Les outils de la **recherche** d'information :
- les **moteurs de recherche sémantiques** (ou linguistiques ou « intelligents ») ;

> les moteurs de recherche non sémantiques (comme **Google** ou **Exalead**) ;
> les « systèmes de question / réponse » sont des moteurs qui permettent de trouver la réponse (et non pas les mots de la question) à une question posée en langage naturel ;
> les outils de **« typologie »** (ou « *clustering* » ou « regroupement » ou « classification automatique ») permettent de catégoriser automatiquement les documents. Ils divisent une base de données en un petit nombre de sous-bases, appelées « classes », telles que deux individus appartenant à une même classe soient aussi semblables que possible et deux individus appartenant à deux classes différentes soient aussi dissemblables que possible. En d'autres termes, la clusterisation tente de décomposer le nuage global de points, représentant l'échantillon, en quelques nuages bien compacts et bien différenciés.

Les outils de **traduction automatique, d'aide à la traduction** et **d'interprétation :**

On distingue deux catégories d'outils de traduction : les outils de traduction automatique et les outils de traduction assistée par ordinateur. Souvent confondues, ces deux technologies relèvent de techniques très différentes, ne visent pas aux mêmes résultats et s'utilisent dans des contextes spécifiques.

> les outils de reconnaissance ou **d'identification** automatique **de la langue** écrite ;
> les outils de **traduction automatique** (TA) (*machine translation,* MT) permettent d'obtenir de façon automatique une traduction (*gist*) de tout type de texte d'une langue (langue source) vers une autre langue (langue cible). Les outils **d'interprétation** (*speech-to-speech translation*) incluent la reconnaissance vocale (ASR), la traduction (*spoken language translation,* SLT) et la synthèse vocale (TTS) ;
> les outils **d'extraction de terminologie** cherchent les termes composés, les expressions-clefs, etc. Le logiciel examine chaque mot du texte source afin de le chercher dans le dictionnaire *ad hoc* et, le cas échéant, proposer automatiquement au traducteur un équivalent cible. L'efficacité de cette fonction est donc essentiellement déterminée par la qualité et par le volume du dictionnaire spécifique ;
> les outils de **constitution de mémoires de traduction** créent des tables d'équivalences (banques de données) entre texte source et texte cible. Pour ce faire, ils divisent le texte à traduire en segments (unités de traduction). Après validation humaine, le logiciel mémorise le segment source et le segment cible comme étant des équivalents linguistiques. Si le segment source apparaît une nouvelle fois dans le texte, le logiciel propose automatiquement la traduction mémorisée. Lors de la mise à jour de la version source d'un texte déjà traduit, le logiciel reprend automatiquement les parties déjà traduites et signale au traducteur les éléments nouveaux ou modifiés. Les logiciels les plus sophistiqués reconnaissent les segments approximativement identiques et les signalent au traducteur en marquant les éléments qui diffèrent du segment mémorisé. Utilisés en réseau, ces logiciels deviennent des outils de travail collaboratifs ;
> l'**alignement** est un processus qui consiste à aligner, c'est-à-dire à poser comme équivalents, segment par segment, un texte source avec le texte cible correspondant.

L'alignement permet de tirer parti de traductions antérieures effectuées sans logiciel d'aide à la traduction en alimentant la mémoire de traduction.

Les outils de **reconnaissance et de synthèse vocale :**

> les outils de **reconnaissance de locuteurs** (qui parle ?) pour l'identification ou la vérification d'identité ;
> les outils de **reconnaissance de la langue** ;
> les outils de **reconnaissance vocale** multilocuteur, multilingue… en milieu bruyant (*automatic speech recognition, spoken language understanding,* SLU) ;
> les outils de **transcription de l'oral vers l'écrit** ;
> les outils de **synthèse vocale** (*text-to-speech,* TTS) ;
> les outils de **« distillation » de la parole** (extraction d'informations) et de **détection des thèmes** ;
> les outils de **détection des entités nommées** ;
> les systèmes de **dialogue** oral (reconnaissance + synthèse + gestion de la conversation).

Les avancées récentes

L'histoire a montré que le développement des technologies est en adéquation avec l'informatisation générale de la société (ordinateur personnel, ordinateurs en réseau, haut débit, etc.) et est totalement connecté à la réalité des usages. Les acteurs principaux sont ceux qui ont porté ces innovations « au bon moment ».

Les technologies de la langue (HLT), terme inventé à la fin des années 90, dénotent une **approche holistique du phénomène de la langue et de son traitement dans toutes ses manifestations, écrites et parlées, combinée avec d'autres modes de communication.** Ce rapprochement même a pu être considéré comme un exploit en cela qu'il rapprochait deux communautés de recherche jusque-là séparées, celle de l'oral et celle de l'écrit, les amenant à échanger sur leurs méthodes, leurs outils, et à travailler ensemble à leur intégration dans des applications utiles. Les objectifs du traitement automatique des langues se sont diversifiés et l'**ambition première de simulation de la compétence langagière est en passe d'être supplantée par un impératif de robustesse** : des applications précises, des environnements réels (bruités, multilocuteurs, etc.), des données textuelles réelles (sur le réseau, les informations télévisées, etc.). Dans ce contexte, on tente de combiner de façon pragmatique les méthodes linguistiques et statistiques [FABRE].

Au cours des dix dernières années, on a pu assister à **la montée en puissance des méthodes statistiques,** dans une logique vertueuse. Nombre de ressources linguistiques, concernant une multitude de langues, annotées à différents niveaux, dont la qualité de traitement est généralement corrélée au degré d'usage de la langue elle-même, ont pu être ainsi constituées. Ces ressources ont rendu possible la fabrication d'outils plus complexes d'indexation, d'analyseurs morpho-syntaxiques, etc. En retour, l'application d'outils statistiques plus

puissants à ces ressources a permis d'opérer des traitements linguistiques beaucoup plus sophistiqués. Cela a aussi facilité le transfert de technologies entre les différentes langues. **À l'heure actuelle, on privilégie la combinaison des méthodes linguistiques et statistiques.** L'articulation des deux démarches prend des formes variées qu'il s'agisse par exemple de filtrer par des mesures statistiques des protocoles déterminés *a priori* ou d'ajouter des probabilités à des règles syntaxiques.

L'arrivée du web sémantique a fait émerger le besoin d'annotation sémantique, à faible coût et à grande vitesse. Les concepts de **l'intelligence artificielle** (AI) comme les ontologies, l'inférence et le raisonnement ont refait leur apparition. La combinaison de ces deux approches scientifiques pourrait rapidement trouver des débouchés applicatifs.

Ces progrès, combinés à l'explosion de l'utilisation de l'internet, ont fait passer les outils de gestion documentaire de la simple recherche de documents à **des fonctions de recherche d'informations, de résumé, de questions / réponses,** etc. Ce passage s'est fait par l'apparition des outils complexes de segmentation, d'extraction de terminologie, etc. Google et ses confrères, malgré de grandes limites en termes de bruit, de silence et de multilinguisme, est devenu un outil indispensable pour l'ensemble des internautes. Néanmoins, il reste de grand progrès à faire concernant la recherche sémantique et multimédia d'informations multilingues.

La qualité de la reconnaissance vocale a énormément augmenté ces dernières années. Des logiciels commerciaux de dictée sont disponibles. La plupart des téléphones mobiles proposent des fonctions de commande vocale. Les centres d'appel ont régulièrement recourt à des commandes vocales. Néanmoins, les outils ne sont encore pas capables aujourd'hui de retranscrire parfaitement des conversations complexes en environnement ouvert.

Malheureusement, au cours des dix dernières années, **le secteur de la traduction n'a que peu profité des progrès** réalisés par ces nouveaux mariages scientifiques. Les progrès les plus remarquables sont plutôt intervenus dans la combinaison entre les méthodes documentaires traditionnelles et les ressources non-écrites comme les images.

Des technologies « diffusantes »

Chaque outil développé n'est généralement pas vendu en tant que produit fini, mais est intégré, en tant que module, à d'autres produits ou services plus complexes (comme le correcteur orthographique dans le logiciel de traitement de texte par exemple) présentés sous forme d'un service rendu à l'utilisateur final.

Les services proposés à partir de l'intégration de technologies linguistiques sont extrêmement dépendants de l'évolution générale de l'informatique. L'éparpillement croissant de la demande, s'il promet des développements nombreux sur le long terme, pose aux entrepri-

ses **un problème majeur d'identification des besoins et des attentes** qui explique probablement une partie de la crise actuelle du secteur. Face à la frustration parfois engendrée par les produits aujourd'hui disponibles, certaines idées reviennent pour toutes les applications : la nécessité de créer des systèmes auto-apprenants, des systèmes capables d'évaluer leur propre niveau d'efficacité (et donc leurs erreurs), des systèmes ergonomiques et des systèmes véritablement multilingues.

La diffusion des ordinateurs personnels connectés, l'expansion des réseaux à haut débit, la convergence des médias requièrent **des outils de recherche toujours plus simples** (pour des utilisateurs toujours plus nombreux et de moins en moins experts en informatique ou en traitement de l'information) et **pour des contenus de plus en plus complexes** (sémantiques, multimédias, multilingues). Ces outils peuvent varier selon l'utilisation, professionnelle ou domestique, qui en est faite.

La multiplication des terminaux personnels, destinés au grand public ou aux professionnels, parfois mobiles (téléphone, télécommande) ou embarqués (voiture), permettant l'échange de contenus multimédias, a requis des interfaces faisant appel au traitement du langage oral et du langage naturel. Des outils équivalents sont aussi nécessaires pour des services tels que les centres d'appel téléphonique multilingues.

Des outils très sophistiqués sont aussi destinés à des professionnels dont le métier est l'information et qui ont donc à traiter systématiquement du langage naturel dans leur chaîne de valorisation de l'information. Il s'agit par exemple de la gestion de documentation multimédia par les documentalistes ou les éditeurs (le catalogue de pièces détachées d'un constructeur automobile, les archives audiovisuelles nationales, les agences de presse), la recherche d'informations stratégiques (le renseignement militaire), économiques ou commerciales, la traduction et l'aide à la traduction (les normes communautaires), etc.

La multiplication des échanges électroniques entre les différentes parties du globe vient créer de nouveaux usages autour du **multilinguisme** (traduction et localisation). Les outils aujourd'hui disponibles sont loin de correspondre aux besoins finançables des institutions et aux besoins sociaux en général.

Quelques grands champs d'application se partagent aujourd'hui l'intégration des Technologies de la langue :

> le « traitement de texte », la « bureautique » ;
> la gestion de documents, structurés ou non : moteurs de recherche et outils de veille ;
> la traduction ;
> les interfaces.

À cette liste, il faut évidemment rajouter **l'industrie des encyclopédies et dictionnaires** qui commercialisent sous forme de produit fini certains de ces développements.

Les moteurs de recherche et les services de veille

Par la **gestion de documents** ou **le traitement de l'information,** on comprend à la fois :

- > la gestion (indexation, recherche, archivage électronique, etc.) d'un patrimoine documentaire ;
- > la gestion de flux informationnels comme la correspondance électronique des grands organismes (les banques, les services publics, etc.) ;
- > les **services de veille** stratégique (*data mining, text mining*) comme la veille sanitaire, l'intelligence économique, le renseignement militaire ;
- > les **moteurs de recherche,** grand public ou professionnels, sur le web ou sur des réseaux fermés ;
- > **les outils d'aide à la décision** ou « d'informatique prévisionnelle » comme les outils de diagnostic médical.

Les moteurs de recherche se focalisent sur deux problèmes : la quantité d'informations à traiter (cf. la taille du web et des index des moteurs) et la capacité à traiter simultanément des milliers de requêtes. Les technologies du « text mining » se focalisent eux sur deux autres problèmes : le traitement d'une grande pluralité de sources et de formats d'informations et les méthodes de classification de l'information.

L'arrivée et la rapidité de développement de **moteurs de recherche** grand public, tels que Google, en moins de 10 ans ont remis la recherche sur la gestion des contenus sous les feux de l'actualité. Que ce soit pour développer un « concurrent » de Google ou créer des services complémentaires, la numérisation croissante de nos activités privées et professionnelles nécessite le développement rapide et efficace d'outils de recherche documentaire, grand public comme professionnels, multimédias, multilingues, sémantiques. Différentes activités de recherche sont mises à contribution, permettant en particulier d'améliorer la chaîne de numérisation, de gérer les formats (conversion, conservation, normalisation), d'améliorer l'ergonomie des terminaux d'accès aux contenus, etc. Toutefois, c'est autour du développement d'outils sémantiques, intelligents, cognitifs d'accès à l'information que l'activité de recherche semble la plus intense. **L'immensité du web et des bases de données propriétaires nécessite de pouvoir interroger des contenus de plus en plus complexes de façon de plus en plus simple** (en langage naturel et dans sa propre langue) ; le développement d'outils d'enrichissement et de classification personnels des contenus venant remettre en cause la vision parcellisée et synchrone sur laquelle reposent encore les outils actuels de la recherche. De grandes activités de recherche comme la création d'ontologie, l'intelligence artificielle se combinent à nouveau avec les technologies de la langue.

Les activités de veille, militaire, économique, sanitaire, etc. sont aujourd'hui, et à court terme, un débouché industriel majeur.

Les outils du multilinguisme et de la traduction

Sous le terme de **« traduction »,** on comprend :

- des outils de reconnaissance de la langue ;
- des **outils d'aide à la traduction** ou de traduction assistée par ordinateur (TAO) qui visent à aider le traducteur, tant au niveau de la cohérence (consistance) de son travail qu'à celui de sa rapidité. Les plus importants de ces outils gèrent, d'une part, la terminologie spécifique du domaine de travail et, d'autre part, des mémoires de traduction. Trados, Déjà Vu, Similis et Transit sont les principaux logiciels de TAO disponibles sur le marché français ;
- des **outils de traduction automatique** (*« to gist »*) comme Systran ou Reverso actuellement en ligne. Les outils d'aide à la traduction disponibles ne sont pas capables de donner des résultats de qualité suffisante pour une publication, mais offrent cependant une aide appréciable lorsqu'il s'agit de comprendre un texte en langue étrangère ;
- des **outils d'interprétation,** généralement encore à l'état de prototypes qui ne fonctionnent que pour une paire de langues dans un contexte très cadré (vocabulaire restreint, locuteur averti...). Ces outils nécessitent d'intégrer un ensemble de technologies complexes, telles la reconnaissance automatique de la parole, la compréhension du langage naturel, la traduction automatique, la génération de langage naturel et la synthèse de la parole.

Depuis les années 1950, la recherche en **traduction automatique** a connu deux grandes méthodes et trois grandes périodes : une période d'approches dites statistiques ou « par apprentissage » ; une période d'approches dites linguistiques ou sémantiques et la période actuelle, depuis la fin des années 1990, plus pragmatique, opère un mixte de ces approches. Pour les tenants des méthodes sémantiques (plutôt les linguistes), il est nécessaire de comprendre un texte pour le traduire (reconnaître la nature et le sens des mots et des liens entre eux) ; pour les tenants des méthodes statistiques (plutôt les informaticiens) cela est inutile puisque la qualité de la traduction repose sur la multiplication d'alignements de corpus annotés par paires de langues. L'algorithme isole des morceaux de phrases, puis recherche dans sa base de données les traductions précédentes de ces groupes de mots, pour choisir enfin la version la plus vraisemblable, sans tenir compte de la syntaxe. Ces méthodes reposent sur l'idée qu'au contraire de l'homme, un logiciel peut analyser la structure d'une phrase, le contexte grammatical, syntaxique, et les transposer dans une autre langue, sans toutefois comprendre le contexte. **Le retour actuel aux approches statistiques s'explique par la croissance de la puissance de calcul des ordinateurs combinée aux capacités de collecte de contenus numériques offertes par l'internet.** L'idée générale est de court-circuiter par les statistiques les difficultés liées à la compréhension automatisée de la syntaxe et de la sémantique, domaine encore trop complexe pour le développement actuel de l'informatique. Ce retour est confirmé par les campagnes d'éva-

luation américaines organisées par le NIST[19]. Au cours de la même période, des progrès notables ont été faits dans des langues peu prises en compte jusqu'alors telles que l'arabe et le mandarin (vers l'anglais).

Les interfaces de communication homme / machine

Sous le terme **« d'interfaces »,** on comprend l'ensemble des applications de dialogue homme / machine en langage naturel (interfaces intelligentes ou cognitives) quels que soient le format (écrit, vocal, multimodal) et le mode d'accès.

La notion d'interfaces peut être appréhendée à plusieurs niveaux :

> le matériel (centres d'appel téléphonique, ordinateur, téléphone, assistants personnels, consoles de jeux, télévision, distributeurs d'argent ou de tickets, robots ménagers...) et les périphériques associés ;
> les logiciels installés (système d'exploitation, de bureautique, d'accès à l'internet...) ;
> les logiciels sur serveurs (les services en ligne).

Concernant les interfaces, il existe des activités de recherche avec des aspects linguistiques et d'autres avec des aspects multimodaux (dont la robotique, la 3D, l'automobile, la téléphonie). **Le développement d'interfaces en langage naturel, écrit et oral, reste un défi majeur des prochaines années ;** principalement du fait de débouchés industriels évidents (les services embarqués dans les voitures, les avions... les centres d'appel téléphonique... les terminaux mobiles comme le téléphone ou les télécommandes, et immobiles comme les distributeurs en tout genre... les jeux vidéo... les systèmes de dictée...). Par ailleurs, la problématique du multilinguisme pour les interfaces reste aussi au cœur de l'innovation. Ce qu'on appelle multilinguisme aujourd'hui ressortit à un processus de « localisation » : adaptation à la langue et à la culture du public visé. Le changement attendu est la prise en compte du multilinguisme, c'est-à-dire de l'adaptabilité à l'infinitude des langues, au cœur des systèmes informatiques. Le passage du standard ASCII au standard Unicode pourrait être l'exemple préhistorique de cette dynamique.

19 Tout récemment, le NIST américain a publié un rapport évaluant les capacités des meilleurs logiciels de traduction automatique à traiter le mandarin et l'arabe. Google a été classé premier. Sur les 40 participants, 8 sont européens dont aucun français et aucun francophone.

Les acteurs

Une estimation économique du secteur semble difficile. Comme toujours dans le secteur culturel ou scientifique, il est très difficile d'estimer la rentabilité d'une innovation car **cette rentabilité est lente et déportée,** particulièrement vers les utilisateurs. Par ailleurs, les technologies de la langue, très « diffusantes », n'ont de sens que comprises entre les activités de recherche et les applications qui les intègrent. Or, **la plupart de ces applications soit n'ont pas de modèle économique stable** (comme les moteurs de recherche grand public), **soit n'ont pas de valeur commerciale évaluable** (comme le renseignement militaire ou la veille sanitaire) ou sont la partie informatique d'un service beaucoup plus vaste et à forte valeur ajoutée humaine (comme l'enseignement des langues ou la traduction). Les 100 millions d'euros (pour la France) ou les 500 millions d'euros (pour l'Europe) de chiffre d'affaire annuel des PME du secteur représenteraient un chiffre aussi irréel que les milliards d'euros obtenus en additionnant le chiffre d'affaires de secteur comme l'enseignement des langues, l'enseignement à distance, la traduction, l'exploitation de quantités massives d'informations multilingues et multimédias, en ligne ou hors ligne, alliée aux techniques de communication hommes / machines, etc.

Avec le Royaume-Uni, l'Allemagne et les Pays-Bas, la France est dans le peloton de tête des industries européennes de la langue. **Il existe en France une longue tradition de recherche, quelques laboratoires et quelques chercheurs de niveau international pour une recherche couvrant à peu près l'ensemble des sujets du secteur.** Certaines PME françaises ont aussi un rayonnement international.

On peut évaluer en France à environ 500 les personnes qualifiées en ingénierie linguistique appartenant tant au secteur de la recherche publique qu'au secteur industriel, dont environ 10% sont des cadres de niveau international.

Les financements publics sont nombreux. Les laboratoires de recherche ont des financements structurels et peuvent répondre aux appels d'offres français ou communautaires spécifiquement réservés aux technologies de la langue. Depuis 2006, les principaux guichets (les programmes de l'ANR en France et le PCRD européen) ne proposent plus d'appui spécifique aux technologies de la langue et cela inquiète les chercheurs du secteur.

Les sociétés privées bénéficient des appuis traditionnels à l'innovation. Elles rencontrent plutôt des difficultés conjoncturelles identiques à celles de nombreuses PME (peu de marchés publics, mondialisation de l'économie...) qui s'ajoutent aux difficultés structurelles du secteur dues aux très importants besoins en financement d'une R&D exigeante.

L'objectif n'est pas ici de recenser de façon exhaustive les acteurs du secteur, mais de baliser le paysage général et de mettre en exergue les principaux acteurs de chaque catégorie et, éventuellement, les relations qui les lient.

Les développeurs d'outils et de services linguistiques informatiques

Une multitude d'acteurs[20] vendent des solutions logicielles déployant diverses fonctionnalités linguistiques. Certaines sociétés sont à la fois éditeurs de logiciels et offreurs de services[21]. **En France**[22], on compte **entre 30 et 150 sociétés** pour **un chiffre d'affaires entre 50 et 100 millions d'euros** selon les sources et les définitions retenues. Près de la moitié des sociétés sont positionnées sur un seul segment d'application. La plupart des sociétés françaises travaillent sur le texte uniquement, mais le segment de la voix est en expansion importante.

La France dispose de **quelques sociétés, hyper-dynamiques, parfois leaders** sur leur créneau. Les PME françaises sont surtout nombreuses sur le secteur de l'écrit. Lingway, Exalead, Sinequa et NewPhénix sont très bien placés dans la recherche d'information multimédia ; Vecsys dans les applications militaires dans le domaine vocal ; Vecsys et Télisma dans les applications pour les centres d'appel téléphonique. Systran, Softissimo et Simos développent des logiciels de traduction. Synaps est l'un des leaders sur la recherche d'information.

Néanmoins, **ces PME sont vulnérables.** Les sociétés **ont des besoins en financement**

importants du fait des coûts de R&D très élevés sur les technologies de base.

Durant les années 1990, de grands groupes français étaient présents sur le marché comme Matra ou France Télécom. Une multitude de nouveaux acteurs a émergé - le nombre de sociétés est passé d'une trentaine à environ 400 - dont plusieurs *spin off*[23] en provenance de centres de recherche publics (New Phénix vient du CEA, Vecsys vient du Limsi, Télisma de France Télécom). La frénésie du marché de l'internet a instauré une course au premier entrant, a développé des partenariats et validé des fusions entre acteurs concurrents. **Depuis les années 2000,** la France a perdu une bonne partie de son tissu industriel.

---

20 Comme dans de nombreux secteurs, pas ou mal identifiés par les nomenclatures d'activité officielles, se posent des problèmes méthodologiques statistiques. Les éditeurs peuvent être des entités juridiques autonomes ou un département d'une société ayant une autre activité principale. La représentativité nationale des sociétés n'est pas toujours facile à identifier. D'une année sur l'autre, le périmètre varie - des sociétés disparaissent (liquidation, rachat ou fusion), d'autres se créent - rendant délicate la comparaison d'une année sur l'autre.

21 ELSNET constitue les annuaires ELSNET / STN des experts et des organisations dans les différents segments du secteur des technologies de la langue et de secteurs connexes tels que celui des centres d'appel téléphonique. L'annuaire recense en France 65 experts appartenant tant au secteur de la recherche publique qu'au secteur industriel (sur 1 170 au total) et 74 « organisations » (sur 2 769 venant de 70 pays). *The Euromap study* [2003] recense 25 laboratoires de recherche et 30 entreprises privées.

22 Au niveau européen, on compte près de 400 sociétés [Bureau Van Dijk, 2005], les cinq premiers acteurs étant Bowne Global Solutions en Irlande, Eckoh Technologies UK, Autonomy UK, Aculab UK, et Scansoft en Belgique. En 2002, le marché était estimé à 510 millions d'euros de chiffre d'affaires, les quatre pays leaders, le Royaume-Uni, la France, l'Allemagne et l'Italie, représentant près de 60% de cette offre. L'offre générale en Europe se répartit entre le traitement du texte (80%) et le traitement de la voix (20%), néanmoins, parmi les cinq leaders, quatre sont spécialisés sur le segment du traitement vocal et la première place du Royaume-Uni peut s'expliquer par la prédominance du traitement vocal dans ce pays.

23 Une *spin-off* est une société nouvelle créée à partir de la scission d'une organisation plus grande, par exemple la formation d'une start-up à partir d'un établissement public de recherche.

Les grands groupes se sont retirés ou ne développent plus que des produits pour les besoins propres de leurs directions commerciales. Les nombreuses fusions et acquisitions au sein du marché peuvent s'expliquer par l'éclatement de la bulle internet, la perte de valeur de plusieurs sociétés et aussi par la pénétration des acteurs américains sur le marché européen, en particulier par le rachat des sociétés qui ont développé des outils de gestion de connaissance dans un domaine hautement stratégique. C'est ainsi que de grosses PME se sont faites racheter par des sociétés qui ont parfois fait faillite (par exemple, la société belge Lernout et Hauspie) ou ont déménagé à l'étranger.

Les marchés sont étroits. De façon générale, **les PME s'épuisent à trouver des débouchés suffisants** sur le long terme. Les outils linguistiques s'intègrent à des services plus larges. Les deux grandes catégories d'intégrateurs sont les fabricants de progiciels (fonction de correction orthographique ou de recherche dans Word par exemple, fonction de traduction automatique dans Google) ou les grands groupes ayant de forts besoins d'outils commercialisables. Or, en Europe, il n'y a pas de grands fabricants de progiciels et donc de grandes marques rassurantes à l'instar de Microsoft, Google ou Nuance. D'autre part, on constate une **inadéquation des produits aux besoins.** Du fait de l'immaturité des produits et du secteur (PME fragiles et absence de normalisation), les intégrateurs potentiels, peinent encore à justifier l'augmentation du prix induit par l'intégration d'un nouveau service (une commande vocale sur une voiture ou un avion par exemple) ainsi que la pérennité de ce service (les mises à jour). Dès lors, **ces PME ne peuvent pas vivre sans les commandes ou les aides publiques.**

Les grands groupes sont en retrait en tant que clients comme en tant que producteurs. Aucun grand groupe européen ne revendique cette activité comme étant stratégique pour lui. Jouve peut être considéré comme une exception. Lorsqu'ils développent des activités de R&D sur les technologies de la langue, c'est en réponse aux besoins de leurs directions commerciales. Il est donc difficile pour les PME de se positionner en sous-traitantes de grands groupes pour répondre à d'importants appels d'offres ou monter de grands projets industriels.

Le secteur est très hétérogène du fait de la multiplicité des approches et des applications et **peu mature.** Rares sont les PME adossées, commercialement ou juridiquement, à un grand groupe. Les dirigeants de ces entreprises sont encore pour la plupart leurs fondateurs et ne sont pas prêts à abandonner leurs prérogatives par des actions de rapprochements industriels. L'association des professionnels des industries de la langue (APIL) est un syndicat professionnel ni très structurant ni très influent.

Du côté de **la recherche,** on retrouve ces caractéristiques de dispersion et de fragilité.

En France, il existe des recherches en linguistique formelle depuis les années 1950, même si le secteur est formellement né au Congrès de Tours de 1982. Ce congrès a donné lieu à une mission interministérielle de développement de l'information scientifique et technique (MIDIST) et au terme « d'industrie de la langue ». Des projets de recherche ont fonctionné

pendant près de 20 ans, principalement autour de la traduction automatique. Vers la fin des années 1990, une déception généralisée s'est installée autour des produits de traduction automatique qui n'ont pas donné les résultats ambitieux escomptés (projet national de traduction assistée par ordinateur en France PNTAO, projet communautaire EUROTRA, projet équivalent au Japon), car la complexité du travail avait été sous-estimée. Cet échec commercial et médiatique ne doit néanmoins pas cacher une réussite scientifique qui peut se mesurer par le nombre et la qualité des chercheurs et des laboratoires que ces projets ont permis d'essaimer. Un échec néanmoins, car il a engendré une frilosité politique quant à la fixation d'objectifs ambitieux sur des sujets comme la traduction automatique, sujet pourtant ancien et porteur de rêves.

On compte **en France une quinzaine de laboratoires publics** dont une poignée a la taille critique pour jouer un rôle international (le LIMSI, l'INRIA si on additionne plusieurs équipes mises bout à bout, LORIA, LADI, TaLaNa, CRIM-INALCO, ENST, CLIPS, GRESEC) et quelques grandes figures individuelles de la recherche.

Probablement pour des raisons culturelles et aussi d'organisation des universités et des grands organismes de recherche, **les laboratoires n'ont pas eu tendance à se regrouper ni à coopérer** véritablement de façon à atteindre une masse critique afin de répondre aux appels d'offres, communautaires par exemple ou à bénéficier de croisement des résultats (entre l'écrit et l'oral ou entre l'informatique et la linguistique par exemple). La volonté de certains laboratoires de travailler uniquement sur la langue française peut apparaître comme un handicap au regard des processus internationaux de financement de la recherche. L'apparition de la recherche sur le multimédia (3D, vidéo, etc.) depuis 2000 environ, a attiré une quantité d'acteurs scientifiques venant de la recherche en technologies de la langue (IBM France et Siemens par exemple ont quasiment cessé leurs recherches en Europe sur cette discipline) sans qu'une nouvelle génération soit encore apparue.

Si l'on peut considérer comme équivalent le niveau de la recherche en Europe et aux États-Unis, il n'en va pas de même concernant l'exploitation industrielle des résultats de la recherche.

Dans ce secteur comme dans beaucoup d'autres, **le transfert des technologies est un point faible.** Toutefois, le fort degré de recherche et d'innovation d'une part et l'immensité des besoins d'autre part rendent cette carence plus évidente qu'ailleurs. Divers facteurs peuvent l'expliquer. La règlementation est très complexe (sur les brevets, les marchés publics, etc.). La normalisation a très peu avancé. Les applications militaires, donc secrètes, sont une partie conséquente des débouchés. Les services achat des grands groupes ne participent pas à une stratégie d'appropriation interne (en « commandant » au service études plutôt qu'en achetant une application commerciale étrangère) ou nationale (en achetant aux PME locales)[24] en privilégiant la sécurité à la qualité. Au niveau européen,

[24] Le *Small Business Act* français ou européen que l'ensemble des PME et de nombreux économistes appellent de leurs vœux ne semblent pas politiquement acceptable par la direction générale de la concurrence de la Commission européenne.

l'hiatus temporel entre la recherche de bon niveau et les besoins commerciaux se retrouvent. L'Europe finance des programmes de recherche pendant que les clients sont déjà en quête de solutions. Les commandes partent vers les industriels non européens qui introduisent alors leurs normes propriétaires rigidifiant ainsi l'avenir.

Les investissements privés sont difficiles à obtenir. Les investisseurs souhaitent des produits commercialisables à moyen terme sans avoir à passer par la construction de ressources et de technologies de base « ingrates ». La coopération entre laboratoires publics et privés n'est pas optimale. La petite taille et le taux élevé de renouvellement des acteurs empêchent les phénomènes de capitalisation et interdisent aux structures françaises de répondre à des appels d'offres, publics ou privés, internationaux.

En conclusion, on peut dire que la situation actuelle est assez tendue. Les technologies développées depuis quelques décennies n'ont pas toujours tenu leurs promesses ni rencontré les besoins des clients (les industriels intégrateurs). Les industriels français sont performants, mais fragiles à cause de leur taille réduite et de leur dépendance à l'égard des grands donneurs d'ordre. Les meilleurs risquent d'être rachetés par des étrangers qui récupèreront leur technologie alors que celle-ci a été largement soutenue par les pouvoirs publics. Le secteur ne dispose pas de grand acteur pouvant servir de lobby ou de moteur à l'échelle industrielle. Du côté de la recherche, malgré quelques personnalités reconnues, peu de laboratoires ont atteint une taille suffisante et bénéficie d'une visibilité internationale. L'arrivée de nouveaux sujets comme le multimédia et la robotique ont eu tendance, ces dernières années, à attirer une partie des financements et des chercheurs du secteur des technologies de la langue.

Les programmes de soutien à la R&D et à l'industrialisation du secteur

Le dossier « technologies de la langue » est principalement traité comme un dossier « recherche ». Au niveau national, le ministère de la Recherche est l'acteur institutionnel principal. Au niveau communautaire, le programme cadre de recherche et développement (PCRD) est l'outil institutionnel dominant. Les dimensions industrielles d'une part et d'usages d'autre part ont peu de relais institutionnels et il n'existe pas vraiment d'instance de coordination de l'ensemble de ces dimensions. Par ailleurs, il n'y a pas non plus toujours la coordination nécessaire entre les trois niveaux de recherche que sont la recherche fondamentale (dont l'extrant est composé de publications scientifiques), la recherche technologique (dont les résultats doivent être régulièrement soumis à des évaluations rigoureuses et impartiales) et les applications de recherche et développement (débouchant sur des prototypes, voire des produits directement commercialisables).

En France, les budgets publics et privés consacrés à la recherche sur les technologies de la langue sont faibles et en diminution, loin derrière ceux d'autres pays ou d'autres langues

comme l'allemand, le néerlandais ou l'italien. Les financements publics actuels ne laissent pas augurer de relèvement de cette situation pour les trois prochaines années. En effet, la fin du programme Technolangue met les acteurs du secteur en position délicate. Concernant le financement de l'innovation industrielle, il existe différents guichets susceptibles d'être ouverts aux sociétés du secteur des technologies de la langue : les pôles de compétitivité régionaux, récemment créés et leur fonds de compétitivité des entreprises géré par le ministère de l'Industrie, OSEO-ANVAR pour les petits projets (1 million d'euros) et l'AII pour les très gros projets (100 millions d'euros). Peut être reste-t-il à inventer un « pôle de compétitivité » national pour les moyens projets (environ 10 millions d'euros) ?

Au niveau européen, le dossier a été pris en charge par des angles divers : les technologies de la langue en tant que telles pendant les années 1980 et 1990, puis leurs applications à partir de la fin des années 1990 (5e et 6e PCRD) : la traduction automatique et les contenus (bibliothèques numériques) pour l'écrit, et les interfaces pour l'oral. Présentement, le dossier est principalement pris en charge au sein du 7e PCRD qui couvre la période 2007 / 2013. Les technologies de la langue sont incluses dans l'axe TIC, l'une des dix priorités du programme, mais n'apparaissent pas clairement malgré l'importante action d'influence menée à l'initiative de la France et de 11 autres États-membres. Il semblerait que l'absence de grands groupes considérant les technologies de la langue comme étant stratégiques pour lui et investissant dans un groupe de pression actif auprès des services de la Commission d'une part, que la vision par les même services d'un paradigme de recherche « à bout de souffle » d'autre part, aient joué un rôle défavorable au cours du travail d'élaboration du programme cadre. Via le 7e PCRD, les projets de recherche continueront d'être financés, probablement *a minima* et en tout cas en concurrence avec les autres disciplines dans l'attente d'une nouvelle « génération scientifique » visible. Par ailleurs, il est de notoriété publique que la réponse aux appels d'offres communautaires par les structures françaises est un exercice difficile, voire dirimant ; *a fortiori* dans des secteurs comme les technologies de la langue où les sociétés sont petites et fragiles, où il n'existe pas de personne morale susceptible d'héberger un consortium, de bureau d'appui, etc.

Ce qu'il faut retenir. Seuls les francophones financeront la recherche, de la plus fondamentale à la plus appliquée, de technologies reposant sur des ressources linguistiques en langue française (unilingues, plurilingues, voire avec le français comme langue pivot). Les besoins en recherche du secteur sont énormes et les industriels, seuls, ne peuvent y faire face.

Les guichets français

> Technolangue et l'ANR

Suivant les conclusions du rapport réalisé sous la présidence de André Danzin pour le Conseil supérieur de la langue française (CSLF) remis au Premier ministre en novembre

2000, il a été décidé la création de Technolangue[25], le premier programme interministériel spécifiquement dédié aux technologies de la langue (écrite et orale).

Ce programme de recherche et développement était doté d'un budget total de 20 millions d'euros, financé à hauteur de 7,5 millions d'euros par les trois ministères de la Recherche, de l'Industrie et de la Culture en articulation avec les trois grands réseaux de recherche et d'innovation technologiques (RRIT) concernés : les télécoms (RNRT), le logiciel (RNTL), l'audiovisuel et le multimédia (RIAM).

Il a démarré en 2002 pour trois ans et a connu une fin effective de l'ensemble des projets financés en 2006. L'appel à projets de 2002 a ouvert un financement à 21 projets dont dix sur les ressources linguistiques, deux sur les standards, un sur la veille et huit campagnes d'évaluation de modules de technologies de la langue ou de systèmes complexes regroupant différents modules.

Il s'agissait d'un programme partenarial associant des laboratoires de recherche publics, des agences, des industriels et des partenaires étrangers apportant leurs propres financements, chaque projet devant être présenté par au moins trois partenaires. L'ensemble des 21 projets a concerné 94 participants différents dont 33 industriels, 39 laboratoires publics, 11 « autres » (des associations, le CEA, la DGA, etc.) et 11 étrangers (Bell Labs / USA, NII / Japon, EPFL, LATL, RALI, etc.). Technolangue a été un outil de promotion du français comme langue pivot avec un grand nombre de langues, notamment des langues minoritaires ayant un faible intérêt commercial, mais un fort intérêt stratégique.

Il y a lieu de s'attarder sur les objectifs de Technolanque et les grands choix de départ, car, **de l'avis unanime, ce programme de R&D était spécialement bien conçu et a généré des résultats conformes aux attentes** compte tenu de sa dotation budgétaire.

Les points forts de Technolangue, à reprendre éventuellement dans un futur programme, sont les suivants :

> l'application d'un principe de subsidiarité par rapport aux autres projets communautaires ou industriels (applicatifs) existants ;
> la priorité donnée à la production de ressources linguistiques ;
> la mutualisation de ces ressources ;
> l'évaluation des technologies et des applications ;
> l'amélioration de la présence française dans les organismes internationaux de standardisation ;
> la veille technologique ;
> le rôle de cohésion professionnelle joué par la réunion régulière du conseil d'administration.

---

25 Voir le site web http://www.technolangue.net.

Le programme Technolangue étant terminé, se pose la question du financement de la recherche française sur les technologies de la langue.

Le problème majeur n'est pas celui du manque d'argent puisque les laboratoires sont financés structurellement et qu'il y a des poches de financement dans les lignes éparses de l'ANR. Il réside plutôt dans le fait que **cette procédure ascendante,** sans un axe directeur préalable et sans justification des choix *a posteriori,* entraîne un évident **manque de visibilité de la thématique** (vis-à-vis du monde de la recherche, des industriels et des instances communautaires) **et l'impossibilité de construire un programme cohérent.**

En **2006,** l'ANR n'a pas lancé de programme spécifique dans ce domaine. Le thème du traitement automatique des langues a néanmoins été inscrit dans la ligne ARA « masses de données et connaissances ambiantes » (MDD). La ligne concerne plutôt la recherche fondamentale et s'adresse uniquement aux laboratoires publics. En 2006, trois projets ont été retenus dans le domaine de l'évaluation (l'équivalent financier d'une première année d'un Technolangue 2). Rien n'est prévu pour **2007.** Il n'est pas écarté qu'il y ait une ligne spécifique en **2008** pour les technologies de la langue si les conditions étaient réunies.

Depuis la création de l'agence nationale de recherche (ANR) en janvier 2006, l'ensemble des décisions de financement des projets de recherche français passe par cette agence. Il n'y a plus de possibilité d'arbitrage interministériel par les services du Premier ministre comme cela s'était fait pour la création de Technolangue. **Les ministères intéressés doivent donc convaincre l'ANR de lancer un programme sur les technologies de la langue.** Ce programme pourra inclure des thématiques insuffisamment prises en compte dans Technolangue 1 ou dans les autres projets actuels comme Quaero : l'image et le multimédia, les interfaces de communication homme / machine, etc.

> Les pôles de compétitivité

Imaginé en 2004 et mis en œuvre depuis 2006, l'outil « pôles de compétitivité » repose sur l'idée que l'innovation est un des facteurs-clefs de compétitivité de l'industrie et qu'elle est d'autant plus efficace qu'elle repose sur des regroupements d'acteurs, dans des entités visibles au plan mondial. L'outil concerne non seulement les domaines technologiques en émergence (nano-technologies, biotechnologies, micro-électronique...), mais également des domaines plus matures (automobile, aéronautique...). Il s'inscrit dans une perspective internationale, en premier lieu européenne. Un pôle de compétitivité résulte de la combinaison, sur un même territoire, de trois ingrédients : des entreprises, des centres de formation et des unités de recherche.

De ces pôles, il est attendu un effet de réseau et une concentration territoriale très bénéfiques bien que certains observateurs craignent que les partenaires des projets se choisissent plus sur des critères de proximité amicale ou géographique (alors que les compétences utiles peuvent être hors de la région) que sur de l'optimisation industrielle ou scientifique. L'objectif est de trouver des moyens (journées de rencontre, consortiums, etc.) **pour**

créer des passerelles entre les mondes très cloisonnés de l'industrie, de la recherche et les grands établissements publics.

Le CIADT du 14 septembre 2004 a souhaité constituer des pôles fondés sur des partenariats publics / privés pouvant impliquer les entreprises, les organismes de recherche et de formation, les établissements financiers, les collectivités territoriales, l'État et l'Union européenne dans le champ des technologies structurantes et des activités industrielles pour lesquelles la France est spécialisée ou bénéficie de potentialités avérées. Suite à un appel à propositions, le CIACT du 6 mars 2006 a labellisé 66 pôles de compétitivité[26], dont 6 projets mondiaux et 10 projets à vocation mondiale. Parmi ceux-ci, certains recouvrent des activités qui ont ou pourraient avoir un lien direct avec les technologies de la langue :

- le Pôle mondial de compétitivité **« Solutions Communicantes Sécurisées » SCS**[27]**,** regroupe les acteurs de la micro-électronique, des logiciels, des télécommunications, des services et usages des TIC de la région PACA ;
- le **Pôle System@Tic**[28] a pour finalité de faire de l'**Île-de-France** l'un des quelques territoires visibles au niveau mondial sur le thème de la conception, de la réalisation et de la maîtrise des systèmes complexes pour quatre marchés applicatifs : télécoms, sécurité/défense, automobile/transports, outils de conception et de développement de systèmes ;
- **Cap Digital**[29] a pour ambition de faire de l'**Île-de-France** l'un des premiers pôles mondiaux des industries du contenu numérique par la mise en place d'une « fertilisation croisée » entre les six domaines stratégiques du Pôle, que sont l'ingénierie des connaissances, le patrimoine, l'éducation, l'image et le son, le jeu vidéo et les services et usages ;
- le **pôle breton Images & Réseaux**[30] est centré sur les nouvelles technologies de l'image et des réseaux (fixes et mobiles) de distribution de contenus. Les projets de recherche, de développement et d'innovation du pôle sont articulés autour de sept axes : les services de la chaîne de l'image, les images en mobilité, les réseaux pour l'image, la distribution électronique de contenus, la sécurité des réseaux, des contenus et des données personnelles, les plates-formes d'acceptance, d'interopérabilité et de convergence, la réalité virtuelle et la réalité augmentée en réseau ;
- **Lyon game**[31] a pour vocation de fédérer les professionnels de la filière en **Rhône-Alpes** du jeu vidéo et des loisirs numériques.

Parmi ceux-ci, le pôle francilien CapDigital semble être celui le plus directement concerné par les technologies de la langue, en particulier le domaine « ingénierie des connaissan-

---

26 Voir www.competitivite.gouv.fr

27 Voir www.pole-scs.org

28 Voir www.systematic-paris-region.org

29 Voir www.capdigital.com

30 Voir www.images-et-reseaux.com

31 Voir http://lyongame.com/accueil/index.php

ces ». Plus avant, il serait imaginable d'envisager de **créer un nouveau domaine « ingénierie linguistique »** afin de favoriser les rapprochements et le montage de projets dans ce domaine spécifique.

Il a été créé un Fonds de compétitivité des entreprises (le FCE) géré par la direction générale des entreprises du ministère de l'Économie. Ce fonds permet d'organiser trois appels à projets par an, non thématisés. Chaque ministère contributeur (industrie, défense, agriculture, équipement) attribue lui-même ses financements. L'enveloppe totale dédiée au financement des pôles, en particulier à leurs projets de R&D, s'élève à un minimum de 1,5 milliard d'euros sur 3 ans.

Pour une PME, la cotisation annuelle à un pôle de compétitivité comme CapDigital est de 250 euros. Elle lui donne accès à tous les services (d'information, de conseils, de rencontres, etc.) du pôle et le droit de participer aux appels à projets. Les projets sont présentés au pôle qui leur octroie, le cas échéant, un petit financement et par là même une sorte de première labellisation, certification de qualité. Ils sont ensuite présentés à d'autres guichets pour obtenir un financement plus important : l'ANR pour un financement par l'un des 3 grands réseaux de recherche, OSEO-ANVAR pour les PME, l'AII pour les très gros projets, un programme communautaire comme le PCRD ou e-Content+, etc.

> AII et le projet Quaero

L'agence pour l'innovation industrielle (AII) a été créée par la loi du 26 juillet 2005 « pour la confiance et la modernisation de l'économie ». Établissement public à caractère industriel et commercial (EPIC), l'AII est placée sous la tutelle du ministre de l'Économie et de l'Industrie. Elle a une dotation de départ de 2 milliards d'euros pour les deux premières années et cherche un modèle économique pérenne pour la suite.

Elle est **la tête de pont d'une politique nationale de grands travaux orientée vers l'innovation.** Elle cherche à maintenir sur le territoire les emplois et la R&D, que les activités soient le fait d'entreprises françaises ou étrangères. Elle a également pour mission d'assurer une veille technologique et une fonction prospective dans le domaine de l'innovation. Elle mobilise les compétences publiques et privées capables de construire des programmes adaptés à la diversité des champs de l'innovation industrielle. L'agence intervient dans un grand nombre de domaines comme la santé, les systèmes d'information et de communication et leurs applications, les transports, le bâtiment, l'énergie et l'environnement, la chimie verte, les biotechnologies. Les projets retenus doivent avoir pour objectif la conception et la réalisation à moyen et long termes de produits impliquant une rupture technologique, lesquels doivent répondre à une demande européenne ou mondiale significative.

L'AII contribue à créer des partenariats capables d'atteindre une dimension mondiale, organisés autour de grandes entreprises ou d'entreprises de taille moyenne, fédérant différents acteurs autour des porteurs de projets, en particulier des établissements publics de recherche et des petites et moyennes entreprises innovantes. Le montage d'un projet pour l'AII

permet à un grand groupe de s'allier à des PME sans passer de commande préalable (moindre prise de risque) en phase de mise au point et sans contrat d'exclusivité léonin au détriment des PME. Toutefois, dans le secteur des industries de la langue, l'absence de grands groupes susceptibles de s'imposer comme un navire amiral naturel pose problème.

L'Agence sélectionne les projets présentés par les industriels, en contrôle la réalisation et fait évaluer régulièrement leur avancement par des experts indépendants. Elle est également chargée du contrôle de ces fonds. Elle peut, le cas échéant, décider l'arrêt d'un programme si celui-ci ne répond plus aux critères définis. Elle peut financer jusqu'à la moitié des dépenses de recherche et développement (R&D) des programmes retenus sous forme de subventions et / ou d'avances remboursables. Ces programmes ont vocation à s'étendre sur une durée de 5 à 10 ans et les financements publics, de l'ordre de 25 à 100 millions d'euros, couvrent une durée moyenne de cinq ans.

L'AII inscrit son action dans un cadre européen. Elle encourage les coopérations transfrontalières au sein de l'espace européen et développe des collaborations et des échanges avec les organismes gouvernementaux semblables des autres États-membres de l'Union européenne. Toutefois, comme la Commission européenne interdit maintenant le cumul des aides nationales et communautaires, il est parfois nécessaire de scinder les activités entre l'innovation et la recherche plus fondamentale qui peut alors être présentée au PCRD. Par ailleurs, il est nécessaire que la direction générale de la concurrence de la Commission européenne donne son approbation au versement des aides nationales et elle est parfois réticente pour les dossiers qui sont les plus proches des débouchés industriels.

Un projet à 100 millions d'euros, sur 10 ans, avec beaucoup d'innovation et donc une grande part de risques, avec des débouchés industriels est un projet type pour l'AII. Avant sa création, aucune structure n'aurait permis d'héberger un projet comme Quaero qui aurait dû se découper en morceaux présentés dans différents guichets.

Déposé dans le cadre de la nouvelle AII, **le programme Quaero**[32] **vise à l'organisation d'une filière industrielle portant sur le développement de nouveaux systèmes de gestion des contenus numériques multimédias** pour des applications grand public ou professionnelles. Il porte sur la conception et la réalisation de solutions nouvelles dans le domaine de la recherche d'informations numériques multimédias et multilingues en se concentrant sur les technologies du traitement automatique de la parole, du langage, de la musique, de l'image et de la vidéo. Quaero n'est donc pas le « Google européen » que certains annonçaient, mais un ensemble de projets de R&D centrés autour de la gestion et de la recherche avancée de contenus multimédias et multilingues.

---

[32] Signifie « je cherche » en latin en tenant compte de la ligature « æ ».

Les développements de Quaero ont vocation à conduire trois grandes catégories d'applications :

> **des outils de recherche multimédias pour le grand public** en environnement résidentiel (ordinateur personnel, télévision, etc.) ou sur téléphone mobile, par exemple pour rechercher des *podcasts*[33], des émissions télévisées, des photos ou des vidéos ;
> **des solutions professionnelles** intégrées de gestion de contenus multimédias, depuis la prise de vue jusqu'à la diffusion, en passant par le montage et la postproduction. Ces solutions permettent une recherche et une gestion efficace et sécurisée des flux audiovisuels dans les chaînes de création, de préparation et de diffusion des contenus ;
> **des solutions de gestion du patrimoine culturel** comme la structuration des archives audiovisuelles et des bibliothèques numériques facilitant leur accès sécurisé pour le grand public et les professionnels des archives.

Les innovations majeures attendues du programme Quaero portent sur :

> l'extension de la capacité de recherche à tous les contenus, y compris les documents radiophoniques, les images fixes, les vidéos et les œuvres musicales ;
> l'automatisation de procédures de génération de descriptifs textuels à partir de contenus multimédias, comme la transcription des bandes son de films ;
> la capacité de naviguer dans de très grandes quantités d'informations ;
> l'extension des mécanismes de recherche et d'accès à l'information aux terminaux grand public tels que les téléphones mobiles ou les téléviseurs ;
> la convergence des procédures de recherche et d'accès à l'information entre environnements grand public et professionnel ;
> l'automatisation des procédures d'adaptation de contenus à différents usages et différents équipements ;
> la prise en compte des contraintes des créateurs et propriétaires de contenus notamment l'interopérabilité des équipements, le stockage et la protection des données.

Le projet est né d'une initiative franco-allemande. Les derniers rebondissements de l'actualité sur le montage économique du projet ne permettent pas de décrire précisément ce montage. Le consortium reste ouvert à de nouveaux partenariats, notamment au niveau européen, dans le cadre d'un processus de cooptation consensuel et rigoureux. Les partenaires actuels sont :

> Thomson France (chef de file du programme) associé à Thomson Grass Valley en Allemagne ;
> de gros industriels : France Télécom et Jouve ;
> des PME innovantes : Exalead, Bertin Technologies, Vecsys, Synapse Développement et LTU Technologies ;
> des laboratoires publics de recherche : le LIMSI-CNRS, qui coordonne les laboratoires publics, l'INRIA, l'IRCAM, l'université Joseph Fourier de Grenoble (CLIPS-IMAG), l'IRIT,

[33] Mot anglais désignant une technique de diffusion de fichiers sonores. Cette technique permet aux utilisateurs de s'inscrire à un flux afin de récupérer automatiquement des fichiers audio. Le terme francisé est baladodiffusion ou diffusion sur baladeur.

l'ENST, le LIPN ou encore le groupe MIG de l'INRA, le RWTH d'Aix-la-Chapelle et l'université de Karlsruhe ;

> des administrations et des EPIC : la délégation générale pour l'armement du ministère de la Défense et le laboratoire national de métrologie et d'essais (LNE) (EPIC rattaché au ministère de l'Industrie), l'INA et la BnF.

Les guichets européens

Comme ses homologues nationaux, **le gouvernement de l'Europe est complexe,** chaque direction reflétant les intérêts de son réseau d'acteurs, la cohérence d'ensemble de ces actions étant, au mieux, un idéal qui s'éloigne au fur et à mesure que l'Union s'agrandit. Les directions concernées par la question des Technologies de la langue sont principalement : la DG Infso (le volet TIC du programme cadre R&D, le programme cadre Compétitivité et croissance et la Bibliothèque numérique européenne), la DG éducation, culture et le nouveau commissaire au multilinguisme, l'agence pour la sécurité (l'ENISA, *European Network and Information Security Agency*).

En outre, si l'on compare ce gouvernement à celui de la France, rajoutons à cette complexité celle du multilinguisme, l'Union européenne reconnaissant le statut de langue officielle à 23 d'entre elles. Cette situation représente un enjeu à deux dimensions : il est, d'une part, nécessaire de préserver les identités linguistiques de chacun des États-membres et des régions dans leur diversité et, d'autre part, de faciliter la communication entre les citoyens de ces différents États. Dès lors, **le dossier des technologies de la langue dans le paysage communautaire apparaît comme double : dans sa dimension industrielle et de recherche comme toutes les autres industries et dans sa dimension politique et culturelle liée au multilinguisme.** Or, il n'existe pas encore de cadre politique communautaire précis pour une Europe multilingue. Il existe une position communautaire sur le multilinguisme, mais il n'y a pas d'unité « Industrie de la langue », ni à la DG Infso ni ailleurs. La direction générale Traduction[34] ne collabore pas au quotidien avec les directions traitant du dossier des technologies de la langue. Ce ne sont pas les même bureaux qui sont chargés du politique (le quoi ? le pourquoi ?) et du pragmatique (le comment ?). Les envolées politiques restent donc trop souvent lettre morte. Un nouveau commissaire en charge du dossier a été nommé au 1er janvier 2007, mais rien n'assure que ses souhaits seront mis en œuvre concomitamment par l'ensemble des directions avec les budgets nécessaires.

Les États-unis ont pris nettement position sur le marché de la recherche et de l'analyse d'information. On connaît le succès commercial de Google, obtenu en moins de 7 ans, et qui lui a conféré une sorte de monopole de fait dans la recherche d'informations

[34] Les instances européennes sont le reflet de ces enjeux. L'Union européenne des 25 comporte 23 langues officielles, soit 506 paires de langues à traduire. La Commission européenne emploie 1650 traducteurs, qui traduisent chaque année plus d'un million de pages. Le Parlement européen emploie 500 traducteurs, consacre un tiers de son budget annuel (soit 300 millions d'euros) à la traduction et l'interprétariat. Malgré cela, les documents officiels mettent du temps à être traduits et beaucoup de réunions se font en anglais, faute d'interprètes, et certains parlementaires européens ne peuvent pas communiquer entre eux, car il n'existe même pas d'interprète ou de traducteur pour certaines paires de langues.

sur le réseau pour le grand public, ce qui représente 1,5 milliard d'internautes naviguant dans un océan de plus de 10 milliards de pages web en 2007. Il ne s'agit que de la partie émergée de l'iceberg : l'État fédéral américain dépense 1 milliard de dollars par an depuis plusieurs années dans la recherche avancée sur les « nouvelles technologies d'analyse de l'information » (NTAI), essentiellement pour les besoins du renseignement et de la lutte anti-terroriste. Par ailleurs, les agences de renseignements (CIA, DIA, NSA, FBI, etc.) se dotent de systèmes avancés de traitement de l'information multi-sources. Enfin, les grandes entreprises du secteur privé bénéficient de ces investissements et répondent en investissant à leur tour : IBM, Microsoft, Google dépensent des centaines de millions de dollars pour développer des plates-formes de traitement intégrées et faire évoluer les moteurs de recherche.

En face de cela, l'idée générale au niveau européen est celle d'**un marché gigantesque en perspective au niveau international et d'un terrain d'expérience inespéré pour les industries européennes de la langue.** La maîtrise de ces technologies **pour répondre au problème du multilinguisme spécifique à l'Europe** placerait nos industriels en position forte pour le traiter dans un cadre mondial. Comme aucun grand groupe industriel européen ne considère ce thème comme stratégique pour lui, c'est en faisant l'analyse et la somme des besoins existants pour chacune des entreprises et de l'intérêt pour les utilisateurs européens, que l'on porte ce thème au niveau de priorité qu'il représente. Il y a donc nécessité d'une action politique pour exprimer et traiter cette priorité. La maîtrise de ces technologies pour répondre au problème du multilinguisme spécifique à l'Europe, placerait donc nos industriels en position forte pour le traiter dans un cadre mondial.

La problématique « recherche » est actuellement l'entrée principale visée par le secteur des technologies de la langue. Les acteurs du secteur cherchent à mettre en œuvre la panoplie la plus large des instruments d'aide à la recherche déployés par la Commission européenne afin d'atteindre une masse critique. Dans ce contexte de partenariats autour de la R&D, le programme Eurêka et le PCRD, bien dotés financièrement, jouent un rôle essentiel en offrant des structures solides pour des projets européens. Ils sont complétés par d'autres partenariats européens, souvent inter-régionaux, dont l'ambition est de créer des synergies et des échanges fructueux pour l'innovation. Parmi les partenariats bilatéraux, le partenariat franco-allemand est stratégique. Enfin, les partenaires européens peuvent s'organiser sous forme d'organisations juridiques européennes.

Il est vrai que **la dimension « linguistique »** propre à chaque zone politique est **fort appropriée à la mise en œuvre du principe de subsidiarité** cher à l'Union européenne, et illustre parfaitement la plus-value que peut apporter une coordination européenne sous l'égide de la Commission. Le nombre de technologies à traiter (recherche d'informations, résumé automatique, reconnaissance et synthèse vocales, dialogue oral homme-machine, traduction automatique du langage écrit et parlé…) et le nombre de langues à couvrir représentent une charge trop importante pour les instances européennes. Cela nécessite et justifie un partenariat entre la Commission et les États-membres. Ainsi, il apparaît normal que les instances communautaires prennent à leur charge la coordination des actions et la recherche d'intérêt général (normalisation, mutualisation des ressources, évaluation) et que

chaque État-membre déploie les efforts nécessaires au développement des ressources linguistiques (corpus, lexiques, dictionnaires...) de quantité et de qualité suffisantes spécifiques à chaque langue.

Ainsi, en juillet 2006 à Luxembourg, François Loos, ministre français délégué à l'Industrie, a présenté les conclusions du groupe de travail du Conseil national de la consommation sur les communications électroniques. La France a préparé une « contribution pour une Europe numérique », visant de fait à accélérer la promotion de l'économie numérique en Europe, inscrite dans la lignée de la stratégie i-2010 menée par la Commission européenne. La proposition contient des recommandations relatives à différents points afférents aux technologies de la langue :

- les réseaux d'information et de connaissances (gestion de grands volumes de masses de données réparties sur le web), la réalisation d'outils de navigation dans ces données (moteurs de recherche, visualisation...), incluant des applications d'intelligence économique ;
- l'interaction personnes/systèmes en développant des interfaces multimodales, incluant la communication vocale ;
- le développement des technologies permettant de traiter les langues européennes individuellement (recherche d'information sur le web, résumé automatique, reconnaissance et synthèse vocales...) ou en termes de traduction, du langage écrit ou parlé.

> Le septième PCRD

Le 7e programme cadre recherche et développement est un programme cadre sur 7 ans (2007 / 2013). Le programme cadre comprend 4 volets spécifiques (la coopération, les personnes, les idées, les « capacités ») et identifie un nombre limité de « challenges » qui collent à des besoins économiques et sociaux pour l'Europe. Il croise de façon matricielle des besoins (bibliothèques numériques, santé, voiture, living et inclusion) et des technologies (réseaux, robotique / cognition, ingénierie / composants).

Les technologies de la langue se retrouvent dans 2 « défis » (challenges) [Commission européenne, 2006][35] » :

- le challenge 2 « Cognitive systems, interaction, robotics » doté de 193 millions d'euros ; l'objectif 2.1 « Cognitive systems, interaction, robotics » (unique objectif du défi 2) ; le sous-objectif 3.2.1.1: « Cognitive Systems, Interaction, Robotics ». Cette ligne porte principalement sur les interfaces et inclut, incidemment, le langage écrit et oral comme l'une des modalités de communication entre l'homme et la machine ;
- le challenge 4 « Digital libraries ». Cette ligne inclut les bibliothèques numériques, le patrimoine culturel immatériel, les « nouveaux contenus », la connaissance « intelligente » et la gestion de la connaissance. Elle comprend donc le traitement automatique de la langue, mais seulement appliqué au contenu.

---

[35] Page 23 du programme de travail

Concrètement, aucun objectif ou sous-objectif entièrement consacré aux technologies de la langue n'a été retenu dans le 7e PCRD, au contraire des deux programmes précédents. Les enjeux liés au multilinguisme et au traitement des langues européennes ne sont que faiblement mentionnés malgré le soutien exprimé par nos représentants nationaux et par ceux de 11 autres États-membres pour ce thème. Ce programme cadre est décliné en programmes spécifiques puis en 7 programmes annuels. Dans le programme de travail 2007 / 2008, doté de 600 millions d'euros, les technologies de la langue n'apparaissent pas plus clairement que dans le programme cadre. Pour la prise en charge de la recherche dans le secteur, il s'agit donc clairement d'un recul, surtout en termes d'affichage.

Côté français, c'est le SGAE qui a compétence pour négocier le programme cadre et les programmes spécifiques. Concernant la négociation des programmes annuels, les représentants français sont des experts du ministère de l'Industrie et de la Recherche. Jusqu'à maintenant, le ministère de la Culture a été très en retrait dans ces négociations.

Le nombre de projets choisis en 2007/2008 dépendra beaucoup du nombre et de la qualité des projets présentés. Or, les projets choisis au titre du PCRD ont un calendrier d'aboutissement commercial à 5/10 ans, délai très long pour les PME fragiles qui constituent presque uniquement ce secteur.

Dans le 7e PCRD, la Commission propose les ERA-Net+ grâce auxquels la Commission finance la seule coordination des programmes nationaux qui continuent eux d'être financés par chacun des États. Une telle proposition déposée en mars 2005, Lang-Net, n'avait pas été retenue même si les technologies de la langue sont un secteur qui semble bien adapté à cet instrument. Il serait loisible de représenter un projet équivalent, porté par les laboratoires publics les plus performants du secteur, où la Commission aurait en priorité la responsabilité d'assurer les fonctions communes (standards, évaluation, communication) et chaque État-membre aurait la responsabilité d'assurer la production de ressources linguistiques pour sa langue et l'adaptation des technologies produites aux spécificités de sa langue.

Une proposition similaire est déjà en cours, l'initiative CLARIN (*Common Language Resources and Technology Infrastructure*), mais elle concerne le développement des ressources linguistiques pour les besoins de la recherche en sciences sociales (en linguistique surtout). Son but est de lutter contre la fragmentation des archives dans ce domaine et de prendre en compte non seulement les langages des pays membres, mais aussi ceux des populations ayant migré vers l'Europe. Le début des opérations est prévu pour 2008 et son coût est estimé à 108 millions d'euros.

> E-Content+

Le nouveau programme e-Content sera intégré au Programme cadre pour l'innovation et la compétitivité (CIP ou PCIC). 2007 est l'année de préparation du nouveau e-Content qui sera exécuté à partir de 2008 (jusqu'à 2020). Les services de la Commission préparent actuel-

lement la constitution d'un réseau thématique visant à recueillir l'information et à réaliser un état des lieux sur les travaux en cours et les ressources linguistiques disponibles afin d'élaborer des recommandations. **À partir de 2009, ce programme pourrait servir à aider la constitution de ressources linguistiques,** le cas échéant.

> Vers un article 169 ?

Il existe depuis de nombreuses années dans le Traité de l'Union européenne un instrument qui peut permettre de mettre en place des actions de plus grande envergure : l'article 169. Celui-ci permet à la Commission d'apporter des financements à un programme conduit en commun par plusieurs États-membres, auquel peuvent participer et contribuer des entreprises. Les crédits ne franchissent pas les frontières, mais sont mis dans un pot virtuel commun, pour financer les partenaires publics et privés qui participent aux projets retenus dans les appels à propositions. La Commission y consacre un budget équivalent à la somme des crédits qu'y mettent les différents États-membres partenaires, avec une plus grande souplesse dans l'affectation transfrontalière des crédits. La part donnée par les entreprises l'est dans le cadre de leur participation au coût total des projets retenus auxquels ils participent, leur financement par les crédits publics étant partiel (30 à 50%). Elles n'ont donc pas à prendre d'engagement préalable.

Cela nécessite cependant une double décision, du Parlement et du Conseil, ce qui fait que très peu d'initiatives ont été tentées. La seule en cours concerne les maladies infectieuses dans les pays tropicaux et ne comporte pas d'activités de R&D.

Les technologies de la langue se prêteraient très naturellement à une coopération entre les États-membres et la Commission, les États ayant en priorité la responsabilité de veiller à ce que les données (corpus, dictionnaires...) existent, de qualité et en quantité suffisante, pour pouvoir développer des technologies utilisables pour leurs langues, et la Commission ayant en priorité le rôle de soutenir la coordination de l'ensemble, la détermination de standards permettant d'échanger les données ainsi que l'évaluation de la qualité des technologies développées afin qu'elle soit suffisante pour couvrir les besoins des applications. Les États-membres et la Commission financeraient en commun les recherches et les développements permettant de produire ces technologies et de réaliser des applications innovantes. Ces applications concernent en particulier la réponse aux besoins propres des activités communautaires : services de traduction, projet de Bibliothèque numérique européenne, service d'alerte proposé par l'ENISA, traduction des brevets communautaires[36]...

Il semble donc que **la nature politique de l'enjeu** lié au traitement du **multilinguisme** en Europe, le développement de technologies appropriées, la composition de l'effort à partager militent pour la préparation d'un tel projet **en parfaite harmonie avec le principe de**

---

36 Le coût important de la traduction obligatoire des brevets est un obstacle au dépôt de brevet. Limiter le nombre de langues dans lesquelles traduire les brevets conduirait sans aucun doute à terme au monopole de l'anglais. Le développement de systèmes de traduction pour ce domaine permettrait de réduire les coûts et d'accélérer le processus, tout en préservant le multilinguisme actuel.

subsidiarité. Cela nécessite une équipe pour s'y consacrer, avec des partenaires étrangers choisis, prêts à s'engager, à soutenir un programme national dans ce domaine et à coopérer[37], un soutien de la Commission (des directions générales Société de l'information, culture, traduction et des commissaires, en particulier du nouveau commissaire en charge du multilinguisme), un soutien du Parlement européen et la mobilisation des industriels européens du domaine.

> Européana, la Bibliothèque numérique européenne

Le 28 avril 2005, six chefs d'État et de gouvernement (Allemagne, Espagne, France, Hongrie, Italie et Pologne), emmenés par la France, ont écrit au président de la Commission européenne afin de lui proposer un projet visant à créer « une bibliothèque numérique européenne, c'est-à-dire une action concertée de mise à disposition large et organisée de notre patrimoine culturel et scientifique sur les réseaux informatiques mondiaux ». Le président Barroso a d'emblée répondu favorablement. Le 24 août 2006 la Commission a adopté une recommandation sur la numérisation et l'accessibilité en ligne du matériel culturel et la conservation numérique.

Le 30 septembre 2005, la Commission a publié, dans le cadre de son initiative « i2010 : une société européenne de l'information »[38], une communication intitulée « i2010 : bibliothèques numériques » qui expose les grandes lignes de l'initiative en la matière et traite de la numérisation, de l'accessibilité en ligne et de la conservation numérique du contenu culturel. Il ressort de la communication que, même si de nombreuses initiatives en faveur de la numérisation ont déjà été prises dans les États-membres, les efforts sont encore dispersés. La numérisation du contenu culturel et l'accessibilité en ligne qui en résulte exigent de relever plusieurs défis. Il s'agit de défis d'ordre économique (qui paiera pour la numérisation ?), organisationnel (comment créer des synergies et éviter les doubles emplois dans les institutions culturelles ? comment assurer la collaboration public / privé ?), technique (comment faire baisser le coût de la numérisation tout en maintenant une qualité élevée ?) et juridique (comment gérer les aspects relatifs aux droits d'auteur ?), sociologique (à qui s'adresser et comment ?).

Cette communication a été bien accueillie par le conseil « Éducation, Jeunesse et Culture » (réunion des ministres européens de la culture) du 14 novembre 2005. Plusieurs ministres ont souligné la nécessité de se fonder sur les initiatives existantes, comme les projets TEL

---

37 12 pays ou régions ont déjà exprimé leur intérêt à participer à une proposition d'ERA-Net dans ce domaine (Allemagne, Danemark, Espagne, France, Italie, Norvège, Suède, République tchèque, Pays-Bas, Belgique, Région du Trentin et Pays Basque), de fortes marques d'intérêt ayant été exprimées par ailleurs par d'autres pays ou régions (Autriche, Finlande, Grèce, Islande, Pologne, Portugal et Régions de Catalogne et de Galice).

38 « i2010 – Une société de l'information pour la croissance et l'emploi », communication de la Commission du 01/06/2005, COM(2005) 229 final. L'initiative i2010 vise à optimiser l'utilisation des technologies de l'information aux fins de la croissance économique, de l'emploi et de la qualité de vie. L'un de ses principaux objectifs politiques est de rendre les contenus numériques européens plus largement accessibles et plus exploitablse pour de nouveaux services et produits d'information.

(The European Library)[39], Michael[40] (qui propose un inventaire européen des collections numérisées), Minerva (Réseau ministériel pour la valorisation des activités de numérisation) et Presto-Space[41] (projet récent qui porte sur la numérisation du patrimoine cinématographique de l'Europe).

Actuellement, le portail TEL mis en œuvre par la CENL (conférence européenne des bibliothèque nationales)[42] constitue une passerelle vers les catalogues des collections des bibliothèques européennes et donne accès à certaines ressources numérisées des bibliothèques participantes. Il s'appuie sur une structure organisationnelle (une équipe de 7 personnes) au sein de laquelle les bibliothèques nationales européennes collaborent déjà et ont acquis une expérience concernant l'amélioration de l'accessibilité en ligne de leurs produits numérisés. Le projet TEL a bénéficié d'un concours financier au titre du 5e PCRD.

À l'issue de la consultation publique organisée fin 2005, la Commission avait annoncé son intention de présenter une proposition de recommandation du Conseil et du Parlement européen, elle a finalement opté pour une recommandation. La présidence finlandaise a donc proposé sur cette base un projet de conclusions qui a été examiné en octobre 2006. Ces conclusions précisent les demandes à la Commission en termes de stimulation, de coordination des travaux, d'évaluation de l'état d'avancement général et d'amélioration du cadre général. **Le calendrier prévoit pour 2008 un accès multilingue aux collections numérisées des bibliothèques nationales à partir du portail TEL et un minimum de 2 millions d'ouvrages numérisés** (livres, images, fichiers sonores, etc.) **et pour 2010 un minimum de 6 millions d'ouvrages numérisés** provenant des collections d'un certain nombre d'archives, de musées et d'autres bibliothèques et, éventuellement, à celles d'éditeurs. La version finale du texte convient aux autorités françaises. Il est à noter qu'une concurrence entre le monde des bibliothèques et celui des archives d'une part, entre le monde l'écrit et celui de l'audiovisuel d'autre part, ne facilite pas toujours l'avancée des dossiers.

La Commission n'entend pas réaliser une base de données unique, mais veut assurer l'accès intégré aux produits numérisés des institutions culturelles de l'Europe. L'utilisateur pourra effectuer des recherches dans différentes collections des institutions culturelles (bibliothèques, archives, musées) à partir d'un seul point d'accès multilingue, qui se présentera sous la forme d'un portail internet. Cela évitera d'être obligé de connaître et de visiter toute une série de sites. Le contenu de la bibliothèque numérique européenne augmentera

---

39 Voir : www.theeuropeanlibrary.org/portal/index.html

40 *Multilingual inventory of cultural heritage in Europe.* Voir : www.michael-culture.org/fr/home

41 *Preservation towards storage and access. Standardised Practices for Audiovisual Contents in Europe.* Voir : www.prestospace.org/index.fr.html

42 La conférence des bibliothécaires nationaux européens (CENL, *Conference of European National Libraries*) vise à accroître et à renforcer le rôle des bibliothèques nationales en Europe, en particulier en ce qui concerne le rôle qu'elles ont à jouer pour assurer la conservation du patrimoine culturel national et l'accessibilité des connaissances dans ce domaine. Les membres de la CENL sont tous les États-membres du Conseil de l'Europe. La CENL comprend actuellement 45 membres de 43 pays européens.

au même rythme que les collections numériques des institutions participantes qui en forment la base. L'initiative relative aux bibliothèques numériques européennes concerne toutes sortes de matériaux : livres, matériel audiovisuel, photographies, documents d'archives, etc. Les archives et les musées seront invités à un stade précoce à apporter leur contribution et à rendre leurs collections accessibles et utilisables par l'intermédiaire de la bibliothèque numérique européenne.

La création d'une bibliothèque virtuelle de dimension européenne nécessitera un effort commun pour avancer sur les trois plans stratégiques suivants :

- numérisation des contenus conservés de manière traditionnelle (textes et photos sur papier, négatifs photographiques, films en rouleaux, musique enregistrée sur disque vinyle ou sur bande, etc.) ;
- accessibilité en ligne de ces contenus ;
- conservation numérique visant à ce que l'information stockée sous forme numérique reste accessible pour les générations futures.

L'engagement des États-membres et des institutions particulières qui participent à ce projet (bibliothèques, archives) sera déterminant pour la rapidité avec laquelle il sera réalisé. Les efforts financiers à consentir pour la numérisation de base nécessaire pour atteindre les premiers objectifs de la bibliothèque numérique européenne peuvent être estimés à quelque 200 / 250 millions d'euros répartis sur une période de quatre ans entre tous les États-membres. Pour l'année 2006, la France a consacré un budget de 3,5 millions d'euros à la modernisation de Gallica, la bibliothèque numérique de la bibliothèque nationale de France, et consacrera un budget de 10 millions d'euros en 2007 à la numérisation de livres du domaine public.

La Commission contribuera dans les domaines où la valeur ajoutée européenne est la plus importante, mais elle ne participera pas au financement de la numérisation de base. Un montant de 60 millions d'euros est prévu dans le cadre du programme e-Content+[43] pour rendre le patrimoine culturel et scientifique de l'Europe plus accessible et utilisable. Un réseau de centres de compétence, cofinancé par la Commission européenne dans le cadre des programmes de recherche, pourrait devenir la pierre angulaire de la numérisation et de la conservation numérique à l'échelle européenne. Ces centres devraient réunir les compétences et les connaissances nécessaires pour atteindre un niveau d'excellence dans les processus de numérisation et de conservation numérique. Ils devront intégrer et exploiter le savoir-faire existant dans les entreprises technologiques, les universités, les institutions culturelles, etc. Les centres de compétence seront choisis à travers des appels à propositions ouverts suivis d'une évaluation des propositions par des experts indépendants.

---

[43] Voir http://cordis.europa.eu/econtent/home.html

Bibliographie

ACCIPIO CONSULTING (Aix-la-Chapelle, Allemagne) pour ITC / IRST (Trento, Italie), *Human language technologies for Europe,* financé par le programme TC-Star (technology and corpora for speech-to-speech translation) de la Commission européenne sous la direction de Gianni Lazzari, http://www.tc-star.org/pubblicazioni/D17_HLT_ENG.pdf, Luxembourg, avril 2006.

AFNOR Standarmedia, la plate-forme des standards et normes des TIC, www.standarmedia.com

ASTIER Hubert, Inspecteur général de l'administration des affaires culturelles, *Rapport d'évaluation de la politique en faveur du français,* ministère de la Culture et de la Communication, Paris, Juin 2005.

BAUDE Olivier, dir., *Corpus oraux, guide des bonnes pratiques,* Presses universitaires d'Orléans, CNRS éditions, Paris, 2006.

BLOCHE Patrick, *Le Désir de France : la présence internationale de la France et de la francophonie dans la société de l'information : rapport au Premier ministre,* La Documentation française, Collection des rapports officiels, Paris, 1999.

BOURBEAU Laurent, PINARD François (Progiciels BPI), *Inventaire des normes clefs pour le traitement informatique du français et adresses des institutions de normalisation et standardisation internationales,* Rapport rendu à RIOFIL (Réseau international des observatoires francophones des industries de la langue), OQIL (Observatoire québécois des industries de la langue), ACCT (Agence de coopération culturelle et technique de la Francophonie), Montréal, Canada, 1995.

Bureau Van Dijk, *Technologies de la langue en Europe : marchés et tendances.* Étude réalisée à la demande du département « Technologie de l'information et de la communication » du ministère français délégué à la Recherche dans le cadre du programme Technolangue, janvier 2005.

COLLET Katell, *Rapport de stage d'analyse documentaire* pour l'URFIST Bretagne Loire-Atlantique, Mars 2003.

Commission européenne, direction générale Société de l'information, *Programme de travail 2007 / 2008 du 7e programme cadre recherche et développement,* Luxembourg, 2007.

Commission européenne, direction générale Société de l'information, *i-2010 : digital libraries,* Office des publications, Luxembourg, 2006.

Commission européenne, direction générale Éducation et culture, http://ec.europa.eu/educa-tion/policies/lang/links/links_en.html, *Liste des institutions et des actions de la Commission européenne dans le domaine de la langue.*

Commission européenne, communication de la Commission au Parlement européen, au Conseil, au Comité économique et social européen et au Comité des régions, *i-2010 : bibliothèques numériques,* Bruxelles, septembre 2005.

Conseil national de la consommation sur les communications électroniques, *Contribution pour une Europe numérique,* conclusions du groupe de travail, présentées par François Loos, ministre délégué à l'Industrie, à la direction générale Société de l'information de la Commission européenne en juillet 2006.

Culture et recherche n° 110, automne 2006, *Culture et programmes cadres de R&D européens : la participation française,* dossier coordonné par la MRT.

DANZIN André, *Rapport du Conseil consultatif sur le traitement informatique,* Paris, 1995.

DEPS / IRI du centre Georges Pompidou / FING, *Compilation des contributions au séminaire Culture 2.0.*

FABRE Cécile, *Traitement automatique de textes : techniques linguistiques,* Université de Toulouse-Le Mirail, équipe de recherche en syntaxe et sémantique (UMR 5610), www.techniques-ingenieur.fr/dossier/traitement_automatique_de_textes_techniques_linguistiques/H7258

FIGEL Ján, REDING Viviane, préface du rapport *« Human language technologies for Europe »,* avril 2006.

GRIIL, Groupe de réflexion sur les industries de l'information et les industries de la langue, *Livre blanc sur « Le traitement automatique des langues dans les industries de l'information »* élaboré par l'association des professionnels de l'industrie de la langue (APIL) et le groupement français de l'industrie de l'information (GFII), janvier 2005.

Intunéo pour le RIAM, *Audiovisuel et multimédia en France : stratégies et priorité de la recherche,* ministère de la Recherche / CNC / ANR, 2007.

JOSCELYN Andrew, LOCKWOOD Rose, *Benchmarking HLT progress in Europe,* The Euromap study, Copenhague, 2003.

LEGENDRE Jean-François, *Plurilinguisme et normes dans les technologies de l'information,* Bilan final 2006 de la convention DGLFLF / AFNOR au titre de l'année 2006, AFNOR département développement.

MORVAN Véronique, Journée d'études de l'ADBS, *Du thésaurus au web sémantique : les langages documentaires ont-il encore un avenir ?,* octobre 2005.

L'actualité langagière, volume 2 / 4, *La traduction automatique : limites et perspectives,* Revue du Bureau de la traduction, ministère des Travaux publics et des Services gouvernementaux, Ottawa, Canada, décembre 2005.

La lettre de la direction générale des entreprises du ministère de l'Économie, *« Dossier : aider les PME à s'inscrire dans l'économie numérique»* n° 18, novembre 2006.

LEVY Maurice, JOUYET Jean-Pierre, *L'économie de l'immatériel. La croissance de demain.* Rapport au ministre de l'Économie, des Finances et de l'Industrie, La Documentation française, Paris, 2007.

MALSEED Mark, VISE David A., *Google Story, enquête sur l'entreprise qui est en train de changer le monde,* Dunod, Paris, 2006.

OUDART Pierre, *« Les politiques publiques en faveur de la diversité culturelle à l'épreuve des territoires »* Culture et recherche n° 106 / 107, décembre 2005.

Science & Technology Forum on Multilingualism, compilation des contributions, Luxembourg, 6 June 2005.

Secrétariat général de la défense nationale, *Comptes-rendus des réunions du groupe de travail sur les outils de la traduction automatique,* documents confidentiels, janvier 2005 à décembre 2006.

Technolangue, le portail des technologies de la langue, www.technolangue.net

TOLILA Paul, *Rapport sur le pilotage de la recherche au ministère de la Culture et de la Communication,* Inspection générale de l'administration des affaires culturelles, Paris, mars 2006.

Lettre de mission

À Madame Jocelyn PIERRE
Ingénieure de recherche

Madame,

De nombreuses initiatives, dans lesquelles le ministère de la Culture et de la Communication joue un rôle parfois prépondérant, sont prises depuis quelque temps dans un ordre relativement dispersé, chacune répondant à une logique qui lui est propre : la Bibliothèque numérique européenne, le projet de guichet unique des bases de données patrimoniales, la Journée européenne des langues, le groupe de travail « intelligence économique » au SGDN, le groupe de travail sur la traduction dans l'administration co-piloté par la DGLFLF et le MINEFI, le démarrage du 7e PCRD et la participation au programme eContent-plus, la place des TIC dans le nouvel organigramme de l'OIF, etc.

Ces initiatives ont cependant en commun de prendre place, plus ou moins directement, dans le domaine du traitement informatisé des langues.

Ce domaine, également connu sous le nom de « traitement automatique des langues », « industries des langues » ou « ingénierie linguistique », vise à fournir des technologies utilisées dans de nombreuses applications qui traitent l'information sous forme écrite ou orale, monolingue ou multilingue, comme les moteurs de recherche, les systèmes de traduction automatique, les logiciels de synthèse vocale, l'analyse de documents, la recherche et le filtrage d'information, le dialogue homme/machine en langage naturel, etc. L'internationalisation des marchés, les capacités de calcul et de mémoire des machines, les avancées en linguistique informatique, la diversité linguistique croissante de la population des internautes, le développement de la téléphonie mobile et maints autres facteurs ont favorisé l'augmentation du volume d'information multilingue conservée et échangée. Dans ce contexte, les aspects d'acquisition, de gestion, d'analyse, d'exploitation, etc. des informations sont au centre des grands débats actuels du monde de la culture, de la recherche et de l'économie.

Dans ces conditions, il paraîtrait utile de disposer d'un « diagnostic réfléchi » sur la question de l'industrie des langues, destiné à cartographier précisément la problématique et à cerner le rôle stratégique que pourrait jouer le ministère.

Ce panorama portera sur les divers volets de la question, considérés globalement :

> un volet « recherche », à traiter en liaison avec la DDAI/MRT, le haut fonctionnaire aux systèmes d'information, le ministère de la Recherche, l'ANR, le CNRS et les Universités, les grands établissements publics culturels tels que l'INA et le CNC, les DG Recherche

et INFSO de la Commission européenne, etc. ;

> un volet « industriel », à traiter en liaison avec le ministère de l'Industrie et les représentants des PME françaises, le service juridique du ministère en ce qui concerne les questions de propriété intellectuelle, les collectivités locales, les instances françaises et internationales de normalisation et de standardisation, etc. ;

> un volet « culturel » à traiter avec l'ensemble des directions du ministère (DLL/CNL, DMF, DAPA), le ministère des Affaires étrangères, les grands établissements publics culturels tels que la BnF, l'Organisation internationale de la francophonie, le Conseil de l'Europe, etc.

Les aspects linguistiques et culturels seront au cœur de cette réflexion. En effet, les langues pour lesquelles des outils performants de traitement automatique ne seront pas disponibles risquent, à terme, d'être exclues des médias modernes de production et de diffusion de l'information professionnelle et non professionnelle. L'existence et l'usage de ces outils renforceront l'usage du français dans tous les domaines et tous les secteurs économiques. Ils joueront aussi un rôle certain dans la vie culturelle, tant en termes de création, de conservation que de circulation des œuvres de l'esprit.

Les conditions me paraissent réunies pour qu'une mission de conseil, auprès de la DGLFLF et avec la collaboration de la DDAI, vous soit confiée. Ses conclusions insisteront sur le rôle spécifique du ministère de la Culture et de la Communication dans l'accompagnement du développement des industries de la langue en vue de la protection et la promotion de la langue française et du multilinguisme ainsi que du développement culturel. Si vous en êtes d'accord, vous pourrez nous remettre un rapport à la mi-janvier, date de la manifestation « Expolangues » à laquelle participera le ministère sur la thématique précise de outils de la traduction.

En vous remerciant de votre disponibilité, je vous prie de croire, Madame, à l'assurance de ma considération distinguée,

Henri PAUL
Directeur du cabinet
du ministre de la Culture et de la Communication

Liste des personnes rencontrées

Mes remerciements pour leur accueil vont à ...

Philippe ALBOU, services du Premier ministre, secrétariat général des affaires européennes, secteur culture

Anila ANGJELI, bibliothèque nationale de France

Fabien ANTOINE, ministère de la Défense, délégation générale pour l'armement, direction de l'expertise technique, département « géographie, imagerie, perception », groupe « traitement d'images et de la parole »

Bruno BACHIMONT, institut national de l'audiovisuel, direction de la recherche

Olivier BAUDE, université d'Orléans, département de linguistique, laboratoire CORAL

Olivier BOSC, ministère de la Culture et de la Communication, cabinet du ministre

Frédéric BOUILLEUX, Organisation internationale de la Francophonie, direction de la culture et de la langue française

Jean CARLIOZ, services du Premier ministre, secrétariat général des affaires européennes, secteur présence française

Roberto CENCIONI, Commission européenne, direction générale Société de l'information, Unité E2

Cyril CHANTRIER, société Temis, membre de l'APIL

Stéphane CHAUDIRON, ministère de la Recherche et de la Technologie, direction de la technologie, département technologies de l'information et de la communication

Khalid CHOUKRI, Association European Language ressources (ELRA / ELDA)

Jean-Michel CORNU, Fondation internet nouvelle génération (FING), direction scientifique

Thierry CRIGNOU, AFNOR, département qualité et production

Jean-Pierre DALBÉRA, ministère de la Culture et de la Communication, musée des civilisations de l'Europe et de la Méditerranée (anciennement ATP)

Victor DAVET, délégation interministérielle à l'aménagement et la compétitivité des territoires DIACT (ex DATAR), mission Pôles de compétitivité

Richard DELMAS, Commission européenne, direction générale Société de l'information

Christophe DESSAUX, ministère de la Culture et la Communication, délégation au développement et aux affaires internationales, mission de la recherche et de la technologie

Emmanuel DÉSVAUX, musée du quai Branly, département recherche, École des hautes études en sciences sociales

Alain DURAND, association française de normalisation (AFNOR)

Jacques FANOUILLAIRE, services du Premier ministre, secrétariat général de la défense nationale

(SGDN), mission pour l'intelligence économique

Christian FLUHR, commissariat à l'énergie atomique, laboratoire d'ingénierie des connaissances multimédias multilangues

Édouard GEOFFROIS, ministère de la Défense, délégation générale pour l'armement, direction de l'expertise technique

Alain GIFFARD, ministère de l'Éducation nationale, de la Recherche et de l'Enseignement supérieur

Didier GIRAUD, institut national de l'audiovisuel, direction de la recherche

Alain LE DIBERDER, société CLVE

Jean-François LEGENDRE, association française de normalisation (AFNOR), direction du développement

Michel LEMONNIER, agence de l'innovation industrielle (AII), programme Systèmes d'information et communications

Olivier LESCURIEUX, institut de recherche et coordination acoustique /musique (IRCAM), relations industrielles

Bénédicte MADINIER, ministère de la Culture et la Communication, délégation à la langue française et aux langues de France

Daniel MALBERT, ministère de la Culture et la Communication, délégation au développement et aux affaires internationales, département des affaires européennes et internationales

Joseph MARIANI, LIMSI-CNRS

Jack MEURISSE, ministère de la Culture et la Communication, secrétariat général, Haut fonctionnaire aux systèmes d'information

Pierre OUDART, ministère de la Culture et la Communication, direction régionale des affaires culturelles d'Île-de-France, service du développement culturel

Pascal POUPET, association française de normalisation (AFNOR), département transport, énergie et communication

Louis POUZIN, European Languages Internet Conference (EUROLINC)

Véronique PROUVOST, ministère de la Culture et la Communication, délégation au développement et aux affaires internationales, mission de la recherche et de la technologie

Michel RABAUD, ministère de la Culture et la Communication, délégation générale à la langue française et aux langues de France

Julie REMFORT, ministère de la Culture et la Communication, délégation générale à la langue française et aux langues de France

Maurice RONAI, École des hautes études en sciences sociales

Jean-Louis ROUVIÈRE, services du Premier ministre, secrétariat général des affaires européennes, secteur TIC

Patrick SCHOULLER, ministère de l'Économie, des Finances et de l'Industrie, direction générale industrie, direction générale des entreprises

Bernard SMITH, Commission européenne, direction générale Société de l'information, Unité E1

Denis SORIOT, ministère des Affaires étrangères, direction générale de la coopération culturelle et du développement, direction de la coopération culturelle et du français, sous-direction du français

Paul TIMMERS, Commission européenne, direction générale Société de l'information, Unité H3 e-inclusion

Laure TOURAINE-PASCAL, services du Premier ministre, secrétariat général des affaires européennes, secteur culture

Xavier TROUSSARD, Commission européenne, direction générale de l'Éducation et de la Culture, direction de la culture, unité culture

Frédéric VIGNAUX, association française de normalisation (AFNOR), département produits et services d'information

Yvo VOLMAN, Commission européenne, direction générale Société de l'information, unité Digital libraries

et à tous ceux sur qui j'ai testé mes idées sans rendez-vous.

Liste des sigles utilisés

ADSL, « Asymetric Digital Subscriver Line » ou « ligne asymétrique numérique ». Il s'agit d'une technologie qui permet de faire transiter à haut débit de très importantes quantités de données numériques sur le réseau téléphonique classique.

AFNOR, association française de normalisation

AFP, agence France-Presse

AII, agence de l'innovation industrielle

ANR, agence nationale de la recherche

ANVAR ou OSEO-ANVAR, agence nationale pour la valorisation de la recherche

APIL, association des professionnels de l'ingénierie de la langue

AUF, agence universitaire de la francophonie

BnF, bibliothèque nationale de France

BNuE, bibliothèque numérique européenne

CEA, commissariat à l'énergie atomique

CENL, *Conference of European National Libraries,* Conférence des bibliothécaires nationaux européens

CIADT ou CIACT, comité interministériel de l'aménagement et du développement du territoire / comité interministériel d'aménagement et de compétitivité des territoires

CISI, comité interministériel pour la société de l'information

CLARIN, *Common language resources and technology infrastructure*

CNC, centre national de la cinématographie

CNL, centre national du livre

CNRS, centre national de la recherche scientifique

CSLF, conseil supérieur de la langue française

CTIL, comité pour le traitement informatique de la langue

DAEI, département des affaires européennes et internationales du ministère de la Culture et la Communication

DAG, direction de l'administration générale du ministère de la Culture et la Communication

DAPA, direction de l'architecture et du patrimoine du ministère de la Culture et la Communication

DARPA, *Defense Advanced Research Projects Agency* (agence américaine en charge des projets en recherche avancée pour la défense)

DDAI, délégation au développement et aux affaires internationales du ministère de la Culture et la Communication

DDM, direction du développement des médias (service du Premier ministre mis à disposition du ministère de la Culture et la Communication)

DEPS, département des études, de la prospective et des statistiques du ministère de la Culture et la Communication

DGA, délégation générale pour l'armement

DG Infso, direction générale Société de l'information et médias de la Commission européenne

DGLFLF, délégation générale à la langue française et aux langues de France

DIACT, délégation interministérielle à l'aménagement et à la compétitivité des territoires (ex DATAR)

DICREAM, dispositif pour la création artistique multimédia

DLL, direction du livre et de la lecture du ministère de la Culture et la Communication

DMF, direction des musées de France du ministère de la Culture et la Communication

DRAC, directions régionales des affaires culturelles du ministère de la Culture et la Communication

DRIRE, directions régionales pour l'industrie et la recherche

EHESS, école des hautes études en sciences sociales

ELDA, *Evaluations and Language resources Distribution Agency,* agence pour l'évaluation et la distribution des ressources linguistiques

ELRA, *European Language Resources Association,* association européenne pour les ressources linguistiques

ELSNET, *European Network of Excellence In Human Language Technologies,* réseau européen dédié au traitement informatique des langues

ENISA, *European network and information security agency,* agence européenne chargée de la sécurité des réseaux et de l'information

EPIC, établissement public à caractère industriel et commercial

FCE, fonds de compétitivité des entreprises

GTN, groupe technique national

IFLA, *International federation of library associations,* fédération internationale des associations de bibliothécaires et d'institutions

INA, institut national de l'audiovisuel

INRIA, institut national de recherche en informatique et en automatique

INTIF, institut francophone des nouvelles technologies de l'information et de la formation de l'OIF dorénavant appelé IFN, institut de la francophonie numérique

ISO, organisation internationale de normalisation

LDC, *linguistic data consortium* (États-Unis d'Amérique)

LNE, laboratoire nationale de métrologie et d'essais

MAE, ministère des Affaires étrangères

MCC, ministère de la Culture et la Communication

MIDIST, mission interministérielle de développement de l'information scientifique et technique

MINEFI, ministère de l'Économie, des Finances et de l'Industrie

MRT, mission recherche et technologie du ministère de la Culture et la Communication

OIF, organisation internationale de la Francophonie

NII, *National Institute of Informatics* (Japon)

NIST, *National Institute of Standards and Technology* (États-Unis d'Amérique)

PCIC, (ou CIP) programme cadre pour l'innovation et la compétitivité de la DG Infso de la Commission européenne

PCRD, programme cadre de recherche et développement de la Commission européenne (ou PCRDT)

RIAM, réseau pour la recherche et l'innovation en audiovisuel et multimédia (créé en 2001 à l'initiative des ministères de la Culture et de la Communication (CNC), de l'Industrie et de la Recherche)

RMN, réunion des musées nationaux

RNTL, réseau pour la recherche et l'innovation en technologies du logiciel

RNRT, réseau pour la recherche et l'innovation en télécommunications

RRIT, réseaux de recherche et d'innovation technologiques

TIC, technologies de l'information et de la communication

SGAE, secrétariat général aux affaires européennes

SGDN, secrétariat général à la défense nationale

Technolangue, action « Technologies de la Langue » (voir chapitre 4)

UIT, union internationale des télécommunications (ITU)

UNESCO, agence des Nations-unies pour l'éducation et la culture

W3C, *world wide web consortium*

WSIS, sommet mondial société de l'information (SMSI)

Remerciements

Je tiens à remercier Henri Paul, Xavier North et Jean d'Haussonville qui m'ont accordé leur confiance en me confiant cette mission ; Jean Sibille pour ses cours quotidiens de rattrapage en linguistique ; Joseph Mariani pour sa disponibilité.

Merci aussi à mes collègues et amis pour leur relecture attentive.

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE
Ministère
Culture
Communication

Délégation générale à la langue française et aux langues de France

6 rue des Pyramides
75001 Paris
téléphone : 01 40 15 73 00
télécopie : 01 40 15 36 76
courriel : dglflf@culture.gouv.fr
www.dglf.culture.gouv.fr