ELDA

European Language Resources Distribution Agency

Agence Européenne de Distribution des Ressources Linguistiques

55-57 rue Brillat Savarin 75013 Paris - France - Tél. : +33 1 43 13 33 33 - Fax : +33 1 43 13 33 30 - Email : choukri@elda.fr

Emilie MARQUOIS

MISE A JOUR DU REPERTOIRE DES OUTILS DE TRAITEMENT AUTOMATIQUE DE LA LANGUE

RAPPORT

Compte rendu de fin de projet d'opération d'une recherche financée par le ministère de l'enseignement supérieur et de la recherche

Date : **15 février 2002** Décision d'aide : **00 K 5202**

Nom de l'organisme scientifique : **ELDA (Evaluation Language resources Distribution Agency - Agence pour l'évaluation et la distribution des ressources linguistiques)**

Nom du responsable scientifique : **CHOUKRI Khalid**

Nom du laboratoire : **ELDA**

Adresse : **55 - 57, rue Brillat Savarin**
75 013 Paris

Numéro d'identification :

ELDA

European Language Resources Distribution Agency

Agence Européenne de Distribution des Ressources Linguistiques

55-57 rue Brillat Savarin 75013 Paris - France - Tél. : +33 1 43 13 33 33 - Fax : +33 1 43 13 33 30 - Email : choukri@elda.fr

RESUME SIGNALETIQUE

Le projet se proposait de réactualiser et d'étendre l'étude réalisée en 1999 pour le Ministère de l'Education Nationale, de la Recherche et de la Technologie, relative à l'inventaire des outils de traitement automatique de la langue (ou outils de TAL). Il se proposait également de diffuser ce répertoire par l'intermédiaire du site Web d'ELDA et de concevoir un outil Internet permettant sa mise à jour régulière et interactive.

Mots-clefs : ingénierie linguistique – outil linguistique

TABLE DES MATIERES

1. Introduction

La société ELDA (Agence pour l'évaluation et la distribution de ressources linguistiques) a été chargée par le Ministère de l'Education Nationale, de la Recherche et de la Technologie, de réactualiser et d'étendre l'étude qu'elle avait réalisée en 1999 pour ce même ministère, relative à l'inventaire des outils de traitement automatique de la langue (ou outils de TAL). Cette mise à jour impliquait le lancement d'une enquête à grande échelle auprès des acteurs de l'ingénierie linguistique, la diffusion des données recueillies par l'intermédiaire du site Web d'ELDA et la conception d'un outil Internet permettant la mise à jour régulière et interactive de ces données.

Ce rapport s'articule de la manière suivante :

Dans un premier temps, nous présentons le contexte dans lequel les projets «Recensement des outils de traitement automatique du français et ressources linguistiques » (1999) et «Mise à jour du répertoire des outils de traitement automatique de la langue » (2001) ont été lancés. Dans un second temps, nous décrivons les différentes étapes mises en œuvre pour réaliser ces enquêtes. Dans un troisième temps, nous analysons les données de la nouvelle enquête et les comparons aux anciennes données. La dernière partie de ce rapport revient sur la mise en ligne des données sur le site Web d'ELDA et sur la conception de l'outil Internet permettant la mise à jour du répertoire.

2. Enquêtes sur les outils de TAL : contexte et intérêts

L'information joue aujourd'hui un rôle de plus en plus important dans toutes les activités du monde contemporain. En effet, de nombreux facteurs, agissant de manière synergique, ont contribué au cours du dernier quart du XX^{ème} siècle, à accroître son influence dans tous les domaines. Simultanément, l'internationalisation des marchés et le développement des Nouvelles Technologies de l'Information et de la Communication (NTIC) ont favorisé le multilinguisme et l'augmentation du volume d'information.

Dans ce contexte, les aspects d'acquisition, de gestion, d'analyse, d'exploitation, etc. des informations, multilingues ou non, pour la plupart sous formes textuelles ou orales, sont au centre des grands débats du monde de la recherche et de l'économie.

L'ingénierie linguistique, confrontation et rapprochement des deux disciplines que sont l'informatique et la linguistique, est ainsi devenue un des domaines-clés qui propose des voies de recherche susceptibles d'apporter des solutions à ces problèmes et répond ainsi aux besoins actuels de ce que les experts dénomment la « société de l'information », les principaux domaines d'applications étant la veille stratégique, la gestion des connaissances et le commerce électronique.

Les outils de TAL sont au centre d'enjeux (essentiellement) économiques (ils peuvent conduire à des gains de productivité dans de nombreux secteurs) et culturels (les langues qui ne seront pas informatisées, c'est-à-dire pour lesquelles des outils performants de traitement automatique ne seront pas disponibles risquent à terme d'être exclues des médias modernes de production et diffusion de l'information, professionnelle ou non). Leur connaissance et leur maîtrise sont essentielles.

Il est ainsi apparu important de tenir compte de l'état de l'art en terme d'approche technique, de donner une description complète et pertinente de ces outils, en vue de leur intégration dans des systèmes complets de traitement de l'information professionnelle, en vue de leur transfert vers des sociétés de développement de technologies. En d'autres termes, leur création, leur identification, leur promotion et leur valorisation dans les milieux industriels, académiques ou autres sont devenues essentielles, voire nécessaires. Mieux soutenir leur développement et leur utilisation l'est également devenu.

De même, prêter une attention à l'évolution du domaine de l'ingénierie des langues et de ses acteurs semble primordial et nécessaire, en ce sens qu'il s'agit d'un secteur porteur qui connaît des bouleversements constants. Il suffit de comparer le paysage des industries de la langue en 1999 à celui d'aujourd'hui pour s'en rendre compte. De nombreux organismes ont soit disparu (c'est le cas de la société Terminotix, spécialisée dans la gestion de terminologie), soit changé de nom (Sinequa a remplacé Cora), soit modifié et/ou réduit leurs activités (c'est le cas de la société Apsydoc qui a réduit ses activités d'ordre multilingue), soit sont apparues (Sémantia, Influo Software, Intelligence Process, etc.) soit ont rencontré ou rencontrent des difficultés. L'exemple le plus marquant est celui fourni par la société belge Lernout & Hauspie.

C'est dans cette optique que de nombreux projets ont été lancés au cours de ces dix dernières années en France, mais aussi en Europe et dans le reste du monde. Il suffit de consulter la liste des projets ayant trait au traitement automatique des langues se trouvant sur le site HLTCentral (http://www.hltcentral.org) pour ce rendre compte à quel point l'ingénierie linguistique prend de plus en plus d'importance aujourd'hui.

Le projet « Recensement des outils de traitement automatique du français et ressources linguistiques » et le présent projet « Mise à jour du répertoire des outils de traitement automatique de la langue » s'inscrivent dans ce mouvement.

3. Mise à jour de l'inventaire des outils de TAL

Nous décrivons ici les différentes étapes mises en œuvre pour réaliser les enquêtes. Il est intéressant de connaître les choix organisationnels et méthodologiques qui ont été faits tout au long de l'enquête, mais aussi lors de la création de la typologie des outils de TAL et du formulaire de description et lors de l'identification des organismes travaillant dans le domaine du TAL et disposant d'outils. Ces choix ont en effet déterminé les informations qui ont été recueillies, informations qui feront l'objet d'une comparaison et d'une analyse dans la quatrième partie. Il nous semble ainsi pertinent de faire référence, même rapidement, à la méthodologie employée.

3.1. Méthodologie employée

Pour mener à bien la première enquête, une liste de types d'outils de traitement automatique des langues et un formulaire de description ont été créés et une liste de prospects, c'est-à-dire d'organismes et/ou de personnes à contacter, constituée[1].

La stratégie employée lors de la seconde enquête était la même que celle définie dans la première. Par ailleurs, la typologie des outils comme le questionnaire n'ont fait l'objet que de légères modifications. Nous n'y ferons donc allusion que très succinctement.

[1] Il faut préciser qu'en complément à cette enquête directe auprès des fournisseurs, ont été ajoutés des outils qui ont été repérés et pour lesquels les organismes concernés n'ont pas répondu.

3.1.1. La typologie des outils de TAL

La première tâche de l'enquête a consisté à définir, cerner et baliser le domaine du TAL, dont les contours sont peu clairs et à définir la notion d'« outil de TAL », les types d'outils et les relations entre ces différents outils.

Le domaine du TAL

On englobe facilement dans le domaine du TAL des outils qui ne comportent pas nécessairement un traitement linguistique. La simple reconnaissance de séquences de caractères n'implique pas nécessairement un traitement linguistique. De même, tous les outils de traitement de l'information ne relèvent pas forcément du domaine du TAL. Si l'on peut compter un grand nombre de moteurs de recherche, notamment sur Internet, ils ne font pas tous appel à des connaissances linguistiques. Seul un examen minutieux de ces outils permet de déterminer s'ils sont concernés ou non par l'enquête.

Des dénominations diverses…

On peut d'ores et déjà attirer l'attention sur le fait que plusieurs termes sont indifféremment employés, y compris par les experts, pour parler de traitement informatisé des langues. Ainsi, le sigle «TAL » renvoie, à lui seul, à plusieurs développements : Traitement Automatique du Langage, Traitement Automatique des Langues, Traitements Automatiques des Langues. On trouve également la variante «TALN », pour Traitement Automatique des Langues Naturelles / du Langage Naturel.

L'expression « Traitement automatique du langage » est largement répandue. Mais, cette dénomination, pour courante qu'elle soit, n'en est pas moins impropre. Elle décalque en effet l'expression anglaise «Natural Language Processing (NLP) » : or, en anglais, le même mot « language » signifie à la fois « langue » et « langage » ; le français, au contraire distingue ces deux termes et comme nous l'avons déjà indiqué, ce sont les langues et non pas le langage qui constituent le support des données linguistiques que les traitements automatiques ont pour objet de manipuler.

Les expressions « Traitement Automatique des Langues Naturelles » et « Traitement Automatique du Langage Naturel » sont également souvent utilisées. L'adjectif « naturel » n'a, dans ces deux cas, qu'une seule et simple fonction : permettre de différencier les langues

dites «naturelles » (c'est-à-dire humaines, telles le français, le chinois, l'anglais, etc.) des langues dites «artificielles » (qui sont les langues que l'homme a tenté de construire, telle l'espéranto, ou bien les langages de programmation).

Nous préférons à dessein utiliser l'expression « Traitement Automatique des Langues (TAL) » (notamment lorsqu'il est question d'outils de TAL) et parler des « traitements automatiques des langues ». En effet, on ne saurait parler « du » traitement mais « des » traitements automatiques des langues ; il en existe une multitude.

Que recouvre cette expression, c'est ce que nous avons cherché à savoir et avons exposé dans le paragraphe qui suit.

Qu'entend-on par « traitements automatiques des langues » ?

L'objectif des traitements automatiques des langues est la conception d'outils capables de traiter de façon automatique des données linguistiques.

Cette définition appelle quelques commentaires sur les mots qui la composent.

Données linguistiques

Les données linguistiques peuvent être de différents types ; il peut s'agir de textes écrits, ou bien de dialogues écrits ou oraux, ou encore d'unités linguistiques de taille inférieure à ce que l'on appelle habituellement des textes (par exemple : des phrases, des énoncés, des groupes de mots ou simplement des mots isolés). La classification des Ressources Linguistiques utilisée par ELRA distingue quatre catégories :

- les ressources orales (lecture de textes à voix haute, discours, dialogues, émissions de radio et de télévision, etc.),
- les ressources écrites (lexiques, corpus textuels, etc.),
- les ressources terminologiques,
- les ressources multimodales/multimédia

Traitement automatique

Traiter, c'est pouvoir analyser, extraire, engendrer, identifier, interpréter, traduire, etc. Le traitement, dit automatique par opposition à un traitement manuel ou instrumental opéré par l'humain, utilise un ordinateur c'est-à-dire une machine conçue pour effectuer des calculs. Un traitement automatique est une suite d'actions ou calculs à faire effectuer par la machine dans un certain ordre chronologique. On parle généralement de programme. Traiter un objet linguistique de façon automatique, implique un certain nombre de contraintes dans la description même de cet objet : il faut pouvoir arriver à formuler de façon totalement explicite et cohérente des ensembles de règles caractérisant le fonctionnement du texte.

Le TAL met en œuvre des outils et des techniques de traitement qui sont de trois ordres : linguistiques, formels, informatiques.

Il faut en effet distinguer la description des connaissances, l'expression de ces connaissances dans un formalisme susceptible d'être implémenté par la machine et l'élaboration de techniques et de stratégies informatiques de traitement effectif.

L'élaboration de systèmes performants passe donc par le détour de recherches fondamentales, en matière notamment de compréhension de texte et de génération de texte. Dans ces deux perspectives, le traitement de la langue porte non seulement sur les formes, mais aussi sur le contenu ; il doit mettre en œuvre des connaissances linguistiques très complètes (relevant des niveaux de la morphologie, de la syntaxe, de la sémantique et de la pragmatique), ainsi que des connaissances d'univers. De telles recherches revêtent nécessairement un caractère pluridisciplinaire, et doivent associer étroitement linguistes et informaticiens.

Définition d'un outil de TAL

Sur la base des éléments précédemment donnés, seront considérés comme appartenant au domaine du TAL, les outils qui ont pour objet des données linguistiques exprimées dans une langue et capables pour les traiter automatiquement :

- d'expliciter les règles de la langue,
- de représenter ces données dans des formalismes opératoires et calculables,
- d'implémenter à l'aide de programmes ces mêmes données.

Il s'agira à la fois d'outils qui traitent la langue et d'outils qui manipulent des connaissances linguistiques pour traiter une masse d'informations.

Ayant délimité le domaine du TAL et défini la notion d'outil de TAL, il nous restait à dresser une liste aussi complète que possible des types d'outils de TAL. Or, les chevauchements entre les applications et les outils d'une part et les interrelations entre les outils eux-mêmes rendaient complexe une « mise à plat » des types.

3.1.2. Choix méthodologiques et organisationnels

Les relations entre les différents outils constituaient la principale difficulté dans l'élaboration d'une typologie simple. Etablir une typologie structurée et basée sur des principes de classification fiables est fort complexe en raison des multiples liens qui existent entre les outils. Plutôt qu'une classification arborescente (de type graphe), nous avons privilégié une présentation linéaire des types d'outils sans préciser les liens qui existent entre eux. Cette démarche a permis de dégager quinze classes principales. Nous retrouvons parmi les grandes classes, les outils d'analyse linguistique, les outils de génération automatique, les outils de traduction automatique et assistée, les outils de traitement automatique de la parole, etc. Chacune de ces classes contient un ensemble d'outils plus précis généralement définis selon la nature du traitement linguistique. A titre d'exemple, on peut distinguer pour la classe des outils de détection et de correction automatique d'erreurs, les correcteurs orthographiques/lexicaux, les correcteurs grammaticaux, les correcteurs stylistique, et d'autres outils plus précis comme les outils de réaccentuation automatique ou les outils d'aide orthographique.

Signalons que certaines sous-catégories d'outils à l'intérieur d'une même classe reposent plutôt sur un besoin méthodologique que sur une distinction empirique. Nous avons, par exemple, détaillé différentes catégories pour les analyseurs ou les correcteurs orthographiques. Or, dans les faits, ces outils combinent en général différents niveaux d'analyse ou de correction. Les outils d'analyse linguistique portent en général sur plusieurs niveaux linguistiques, seul le degré de complexité des traitements varie. De façon identique, un correcteur se limite rarement à une correction purement lexicale. Il en va de même pour les autres outils qui peuvent appartenir à différents types. Certains outils combinent ainsi par exemple des fonctions de recherche et d'indexation. Dans des cas comme celui-ci, un type a

été privilégié au dépend de l'autre. Le choix fait est un choix arbitraire qui méritait une analyse poussée.

Nous avons également répertorié quelques types d'outils dont le caractère de TAL est moins apparent mais qui sont souvent intégrés dans des outils de traitement de la langue. C'est notamment le cas pour des outils de statistiques lexicales[2] ou de segmentation[3] utilisés dans le traitement de documents. Nous avons estimé que le recensement de tels outils était important car il s'agit souvent de composantes d'outils plus complexes et on constate que, de plus en plus, des modules de TAL, comme des analyseurs syntaxiques de surface ou des ressources linguistiques, les intègrent.

La typologie des outils de TAL (Annexe 1) qui a été utilisée dans le cadre de la seconde enquête a subi de légères modifications. Nous avons cherché à l'alléger et tenté d'uniformiser les formulations utilisées. Certains types d'outils ont été supprimés soit parce qu'aucun outil n'avait été identifié pour ces catégories lors de la première enquête (« prédicteur de traduction » et « analyse de contenu ») soit parce qu'ils étaient récurrents (« Système de vérification automatique de traduction » vs « outil de post-édition automatique » et « Interfaces en langage naturel » vs « interrogation en langage naturel »). Un seul type a été ajouté, dans la catégorie « Aide à la rédaction », intitulé « Autres », qui a pour but de rassembler tous les outils appartenant à cette catégorie mais n'appartenant pas aux types identifiés. La partie qui a finalement été la plus modifiée est la partie « Ressources linguistiques ». Cette partie est désormais organisée selon le schéma oral/écrit/multimodal/multimédia/grammaires/autres.

Une fois la liste des types d'outils de TAL établie, nous avions à répertorier les outils existants dans les milieux de la recherche et de l'industrie. La réalisation de formulaires de description pour ces outils et l'établissement d'une liste de producteurs d'outils constituaient donc les étapes suivantes de notre enquête.

---

[2] Les outils de statistiques lexicales calculent la fréquence d'apparition de termes et sont souvent à la base d'outils d'indexation automatique.

[3] Les outils de segmentation permettent de segmenter toutes productions orales ou écrites selon différents niveaux : un texte peut être segmenté en paragraphes, phrases, mots, caractères. Les outils de segmentation sont par exemple utilisés par les analyseurs, les aligneurs de corpus, les systèmes de reconnaissance de caractères, etc.

3.2. Le questionnaire

La conception du questionnaire devait répondre à plusieurs impératifs afin de permettre d'obtenir les caractéristiques essentielles d'un outil (type de l'outil, plates-formes matérielle et logicielle requises, aspects juridiques et commerciaux, intégrations existantes ou envisageables, modèle linguistique adopté, démarches linguistique et informatique, etc.). Nous avons fait le choix de détailler et d'homogénéiser les critères portant sur les aspects techniques et pratiques et de laisser une partie libre pour la description « scientifique ».

Le questionnaire est ainsi constitué de quatre parties :

La première partie concerne l'identification de l'organisme fournisseur et les partenariats éventuels.

La deuxième partie du questionnaire porte sur l'identification de l'outil et comporte les rubriques suivantes : nom de l'outil, le type, son utilisation et sa maturité (disponibilité, utilisateurs potentiels, etc.).

La troisième partie du questionnaire porte sur l'aspect technique (support, taille des données, plate-forme matérielle et logicielle, etc.) et commercial (disponibilité d'une documentation, contraintes à la commercialisation, coordonnées du distributeur, etc.).

La quatrième est une zone libre destinée à la description détaillée des aspects linguistiques et informatiques. Elle comprends également une zone de commentaires (informations connexes, bibliographie éventuelle, outils ou ressources développés par des partenaires, etc.).

L'adoption de critères de description fixes et prédéfinis pour les aspects linguistiques et informatiques aurait, certes, permis d'obtenir des réponses homogènes et des points de comparaison. Nous n'avons pas retenu une présentation par rubriques prédéfinies, comme nous l'avions fait pour les autres critères, pour plusieurs raisons. Tout d'abord, une telle présentation aurait considérablement alourdi le questionnaire ; or, il s'agissait pour nous de proposer un questionnaire assez court et facile à compléter afin de favoriser le taux de retour. Par ailleurs, cette présentation, plutôt contraignante nous semble-t-il, risquait d'appauvrir la richesse des réponses. Nous l'avons constaté pour d'autres critères, plus simples, comme celui

de la disponibilité ou de l'existence d'une documentation lorsque le prospect voulait donner une information plus précise qu'une réponse affirmative ou négative, ou une simple date. Nous avons estimé que ce choix permettait aux organismes de fournir les informations souhaitées, et leur facilitait éventuellement le travail puisqu'il permettait d'intégrer une description déjà existante.

Dans le cadre de la seconde enquête, le questionnaire (Annexe 2) devait répondre aux mêmes impératifs que ceux de la première enquête afin que les réponses obtenues soient homogènes et comparables. Il nous a néanmoins semblé qu'envoyer un questionnaire légèrement différent pouvait être un atout. Cela permettait de montrer que nous avions réfléchi à cette nouvelle enquête. Par ailleurs, le domaine de l'ingénierie linguistique ayant évolué, le questionnaire se devait d'avoir évolué aussi. Enfin, l'enquête précédente avait permis de voir que le questionnaire n'était pas suffisamment structuré. Nous lui avons donc apporté quelques modifications qui consistent essentiellement dans la réorganisation des divers champs à remplir : ajout de champs (« Raison sociale », « Type », « Rue » et « Ville »), changement du nom de certains champs (par exemple, le terme « Toile » a été remplacé par l'expression «Site Web »), déplacement de champs (par exemple, le champs « Nom » a été placé en tête et remplace de ce fait le champs « S'agit-il de »), etc.

3.3. La liste des prospects

Le travail de repérage des fournisseurs potentiels d'outils a été essentiellement réalisé lors de la première enquête. Il a permis d'établir une liste de 253 organismes (pour lesquels nous disposions d'un minimum d'informations[4]), qui a par ailleurs été utilisée, modifiée et complétée lors de la seconde enquête.

Nous avons choisi de ne conserver que les organismes développant et / ou distribuant un(des) outils) de traitement automatique de la langue. C'est la raison pour laquelle, avant de contacter les organismes, nous nous sommes assurés qu'ils étaient toujours actifs et qu'ils proposaient ce type d'outils. Une sélection a ainsi été réalisée et des organismes comme l'Aérospatiale, Adobe Systems France, l'AFL (Association Française pour la Lecture) ou encore la SNCF ont été retirés du répertoire. Précisons par ailleurs que les informations

[4] Les organismes pour lesquels nous avions des informations insuffisantes ou erronées (changement de nom, intégration dans un autre groupe, adresse électronique incorrecte etc.) ont été mis de côté pour le moment.

obtenues étant présentées différemment, des regroupements et des distinctions ont été effectués. Nous avons distingué dans certains cas, l'université, le groupe / département de recherche, l'équipe de recherche et la personne à contacter. Pour chacun de ces ensembles, nous avons saisi des coordonnées lorsque c'était possible.

Les principales sources qui ont été utilisées pour le repérage lors de ces deux enquêtes sont citées ci-dessous.

1) La base de données d'ELDA qui comporte plus de 1200 contacts dont près de 300 en France ;

2) Les adresses d'organismes du domaine de l'ingénierie linguistique figurant sur le site de la DGLF (Délégation Générale pour la Langue Française)[5]. Indiquons que dans le cadre du second projet, un accord de mise en commun des données figurant sur ce site et des données dont disposait ELDA a été conclu ;

3) Divers annuaires, répertoires, guides thématiques ou spécialisés ont été fort utiles. Signalons entre autres :
- *Le Répertoire des organismes membres de l'Aupelf-Uref* [Mariani J., 1998] qui recense des centres de recherche francophones et quelques organismes privés ;
- *The Language Engineering Directory* [Hearn P. M.,1996] qui recense les organismes en ingénierie linguistique et les produits qu'ils développent. Cet annuaire a permis de recueillir des informations sur de nombreux outils (nom, catégorie, fournisseur et description sommaire), mais de nombreuses informations étaient malheureusement déjà obsolètes en 1999 ;
- *The Telematics Applications Programme in Language Engineering*-1994-98, réalisé par la DGXIII de la Commission Européenne, a permis de repérer les organismes français participant à des projets européens ;
- Divers guides d'expositions sur les langues (*Catalogue Expolangues*-1999), les nouvelles technologies, la GED, etc. (*Catalogue IDT-97)* ;

[5] URL : http://mistral.culture.fr/culture/dglf/riofil/menu.htm

4) Les actes de colloques ont également permis de repérer des organismes et des outils de TAL. Ils sont une véritable mine d'informations, la thématique étant bien définie et les intervenants nombreux. Citons, par exemple, les actes des premières *JST Francil-1997*, les actes d'*LREC 1998* et d'*LREC 2000*, les actes *JEP, XXIIe Journées d'Etudes sur la Parole-1998*, les actes du colloque ISKO 2001 sur le résumé et le filtrage, etc.

5) Diverses revues spécialisées, telles que *La lettre d'information d'ELRA*, la revue *T.A.L.*, *La lettre d'information de FRANCIL, Langages, Langue Française, L'Information grammaticale*, etc.

6) D'autres documents scientifiques tels que les ouvrages *Fondements et perspectives en traitement automatique de la parole* [Meloni H., 1996] et *Linguistique et Traitements Automatiques des Langues* [Fuchs, C., 1993].

7) Le Web a également constitué une importante source d'information. Nous avons consulté les sites déjà repérés et avons effectué des requêtes à partir des moteurs de recherche. Pour nos requêtes, nous avons privilégié deux moteurs de recherche : le moteur de recherche *Voilà*, qui, nous a semblé assez efficace et qui par ailleurs effectue la majorité de ses recherches sur des sites francophones. Nous avons aussi utilisé *Alta Vista* qui, malgré les redondances et les bruits, permet de repérer des informations pertinentes. Ce dernier nous a permis notamment d'élargir notre investigation sur la plan géographique (Canada, USA, etc.) et ainsi de repérer d'autres outils portant sur le français et disponibles en France.

L'examen de ces diverses sources a permis de recueillir des informations sur les coordonnées de la personne à contacter, les outils disponibles, le profil de l'organisme, etc., et ainsi de faire une prospection directe et aussi ciblée que possible.

Toutes ces informations nécessitaient d'être vérifiées, complétées et mises à jour par une prospection directe car tous ces matériaux présentent des limites[6].

[6] Les guides sont vite obsolètes car les organismes s'intègrent à d'autres groupes, changent de nom, ou n'existent plus, les outils évoluent, s'intègrent dans d'autres outils.
Les articles scientifiques font le point sur l'état d'avancement d'un produit, ces descriptions sont souvent pointues et elles mentionnent rarement les applications industrielles envisageable pour un outil, à l'inverse, les fiches-produits (de GED notamment) du secteur privé portent plutôt sur les fonctionnalités d'un outil et moins sur les aspects linguistique et informatique.

3.4. Déroulement de l'enquête

Une fois la liste de prospects établie, nous leur avons adressé individuellement un courrier et un formulaire de description.

La solution de mettre le questionnaire sur le site Web et d'attendre les réponses n'a pas été retenue dans la mesure où cette démarche «un peu passive » aurait demandé davantage de temps. Par ailleurs, nous ne l'avons pas diffusé via des listes électroniques. Nous avons privilégié une enquête directe auprès des acteurs de l'ingénierie linguistique.

La proposition de pré-remplir au maximum les formulaires étaient généralement bien accueillie par nos interlocuteurs, la raison essentielle invoquée étant le gain de temps. Pour la première enquête, nous n'avons complété que la partie comportant les coordonnées de l'organisme sans mentionner, dans un premier temps, l'outil qui nous intéressait afin de ne pas limiter la réponse de notre interlocuteur. Comme l'a montré par la suite notre enquête, certains organismes qui disposaient de plusieurs outils de traitement automatique ont fait le choix de n'en décrire que certains ou bien nous ont fourni des informations plus récentes que celles que nous avions pu recueillir. Pour la seconde enquête, nous avons envoyé les questionnaires comprenant les informations que les prospects nous avaient déjà fournies. Seule la partie destinée à la description détaillée de l'outil ou de la ressource linguistique était différente puisque nous avions réalisé un résumé à partir de l'ensemble des informations fournies. Pour les nouveaux contacts, nous avons procédés de la même manière que lors de la première enquête : nous avons complété la zone de coordonnées des questionnaires.

Le choix d'un format électronique standard pour le questionnaire (RTF) a été généralement bien accueilli, nous l'avons néanmoins adressé sous d'autres formes à certains prospects sur leur demande (courrier, texte et Word).

Les sites Web des organismes de recherche, bien qu'indispensables, ne sont pas toujours très clairs, complets et les mises à jour ne sont qu'exceptionnellement récentes. De sorte que de simples recherches comme : quels sont les laboratoires ou équipes qui font du TAL, qui en est le directeur, quels sont les thèmes de recherche, quels sont les logiciels réalisés ou en cours de réalisation ? etc., supposent chez l'utilisateur une certaine persévérance. Il faut reconnaître que souvent, ce sont les programmes de séminaires ou de colloques, ou mieux encore les résumés des interventions qui nous ont mis « sur la piste d'un outil ». Les sites des organismes privés, bien que plus professionnels et d'aspect plus commercial vont « droit au but », il suffit de cliquer sur un bouton « produits » pour en avoir la liste. Ceci étant, l'information recherchée n'y figure pas toujours, et il est souvent difficile d'évaluer à partir d'une description Web si l'outil en question concerne ou non notre enquête.

La prospection s'est faite pour les deux enquêtes en plusieurs étapes (Tableau 1). Dans un premier temps, nous avons adressé un(des) questionnaire(s) ainsi qu'un courrier expliquant notre démarche. Une fois la date de retour des questionnaires dépassée, nous avons fait une première relance. Nous avons fait un deuxième courrier individualisé et joint une nouvelle fois le(s) questionnaire(s) pré-rempli(s). Environ un mois plus tard, nous avons envoyé un troisième courrier assez court et non personnalisé annonçant la clôture de l'enquête aux prospects n'ayant pas répondu. Plusieurs jours après cette troisième relance, nous avons contacté par téléphone les acteurs qui n'avaient pas répondu à l'enquête. Certains nous ont envoyé le(s) questionnaire(s).

Tableau 1 : Agenda des deux enquêtes

	Enquête 1 (1999)	**Enquête 2 (2001)**
Envoi des premiers courriers	mi-novembre	début juillet
Relance n° 1	mi-décembre - début janvier	mi-juillet - fin juillet
Relance n° 2	mi-février	début septembre
Appels téléphoniques	fin février	mi-septembre

4. Les données

Combien d'organismes et de personnes ont été contactés pour l'enquête ?, combien d'organismes ont renvoyé les formulaires de description ?, combien de formulaires ont été mis à jour, combien d'outils et de ressources ont été identifiées ?, etc. C'est ce que nous allons voir dans le paragraphe 4.1.

Dans le paragraphe suivant (4.2.), nous revenons sur les données obtenues.

4.1. Les réponses reçues et les données recueillies par ELDA

Avant de présenter les résultats obtenus lors de l'enquête, soulignons que le nombre de questionnaires remplis et retournés par les prospects sont relativement faibles.

Lors de la première enquête, sur 253 organismes prospectés, seuls 83 organismes ont répondu en nous retournant un (ou plusieurs) questionnaire(s) (soit ~ 33 % de retour). Cela

nous a permis de recueillir 161 questionnaires décrivant des outils (143) et ressources linguistiques (18).

Lors de la seconde enquête, sur 139 organismes contactés (pour 190 personnes contactées), 62 organismes ont répondu (soit ~ 45 % de retour), soit en nous retournant un (ou plusieurs) questionnaire(s) mis à jour, soit en nous indiquant qu'il n'était pas nécessaire de mettre à jour le(s) questionnaire(s), soit enfin en nous priant de retirer tel ou tel outil du catalogue, parce qu'obsolète, plus développé ou plus distribué, soit en nous signalant l'existence d'un ou plusieurs nouveaux outils qui n'avaient pas été répertoriés lors de la première enquête parce que nouvellement créés. En complément à la prospection directe, nous avons répertorié un certain nombre d'outils soit parce que les organismes prospectés n'avaient pas répondu, soit parce que les informations collectées étaient insuffisantes. Nous avons également répertorié des outils d'organismes qui n'avaient pas été prospectés lorsque le « hasard de la navigation Web » nous y autorisait. Les Tableaux 2 et 3 fournissent les chiffres exacts concernant ces différents cas de figure.

Tableau 2 : Cas de figure pour les outils identifiés lors de la 1ère enquête

Cas de figure	Organismes	ELDA	TOTAL
Outils à retirer	61	25	86
Outils mis à jour	76	47	123
Outils sans modification	21	26	47
Outils à mettre à jour	-	-	27
TOTAL	158	98	283

26 outils doivent encore être mis à jour. C'est ce qui explique que dans la case TOTAL on trouve la somme 256.

Tableau 3 : Cas de figure pour les outils nouvellement identifiés

Cas de figure	TOTAL
Nouveaux outils signalés par des organismes se trouvant dans la base à l'issue de la 1ère enquête	24
Nouveaux outils identifiés par ELDA d'organismes se trouvant dans la base à l'issue de la 1ère enquête et non signalés par eux	15
Nouveaux outils d'organismes identifiés par ELDA ne se trouvant pas dans la base à l'issue de la 1ère enquête	52
TOTAL	91

Les 139 organismes contactés sont répartis selon les catégories organisme public, université et entreprise (Tableau 4) et sont rattachés pour certains au CNRS (Tableau 5).

Tableau 4 : Type d'organismes

Type d'organisme	TOTAL
Organisme public	13
Université	31
Entreprise	95
TOTAL	139

Tableau 5 : Organismes rattachés au CNRS

Type d'organisme	Organismes rattachés au CNRS
Organisme public	1
Université	19
TOTAL	20

19 organismes de recherche se trouvant au sein d'une université sont rattachés au CNRS. Un seul organisme public a été identifié. Il s'agit de l'INIST.

Le recueil de toutes ces informations (à partir des formulaires et à partir de nos diverses sources) a permis de constituer l'inventaire définitif des outils. Nous avons voulu faire un inventaire aussi exhaustif que possible, cependant certains éléments ont certainement échappé à notre vigilance. La coopération des acteurs du domaine et la disponibilité des informations, notamment dans le secteur de la recherche, ont compliqué le travail de recensement.

4.2. Les outils recensés

Cette partie se propose de faire un bilan concernant les données obtenues, notamment en termes de types d'outils. Nous ferons également rapidement le point sur les caractéristiques suivantes : date de disponibilité et de commercialisation, maturité, disponibilité juridique, commerciale et technique, transferts de technologie, types d'utilisateurs, plate-forme standard et langues traitées.

4.2.1. Typologie des outils recensés

Le Tableau 6 récapitule les outils selon leur domaine d'appartenance pour l'enquête 1 et pour l'enquête 2 (cellules en italique et grisées).

Tableau 6 : Nombre d'outils par type

Type d'outils	**TOTAL**
1. Ressources Linguistiques	22
1. Ressources Linguistiques	*26*
2. Analyseurs Linguistiques	28
2. Outils d'analyse linguistique	*26*
3. Outils de Génération Automatique	9
3. Outils de génération automatique	*2*
4. Outils de Traduction Automatique et aide à la traduction	22
4. Outils de traduction automatique assistée	*21*
5. Systèmes de Résumé Automatique	4

5. Outils de résumé automatique	*3*
6. Systèmes de Compréhension du LN	3
6. Outils de compréhension du langage naturel	*0*
7. Systèmes de Gestion de Terminologie	20
7. Outils de gestion de terminologie	*12*
8. Outils de Traitement Automatique de la Parole	41
8. Outils de traitement automatique de la parole	*52*
9. Outils de GED et Recherche d'informations	45
9. Outils de gestion électronique de documents et recherche d'informations	*52*
10. Outils d'Aide à la Rédaction / aux auteurs	17
10. Outils d'aide à la rédaction	*16*
11. Outils de Traitement Optique de Caractères	9
11. Outils de traitement optique de caractères	*7*
12. Outils d'Apprentissage assisté par ordinateur	38
12. Outils d'enseignement intelligemment assisté par ordinateur (EIAO)	*38*
13. Outils d'Aide à la construction d'outils ou de ressources	18
13. Outils d'aide à la construction d'outils ou de ressources	*16*
14. Outils d'évaluation de systèmes de TALN	1
14. Outils d'évaluation de systèmes de TAL	*0*
15. Autres outils	6
15. Autres outils	*12*
TOTAL	282
TOTAL	*283*

De manière générale, le nombre des outils pour chaque type est peu différent de celui qui avait été trouvé lors de la première enquête. Les outils de traitement automatique de la parole et de gestion électronique de documents constituent toujours la majorité des outils identifiés

avec chacun un total de 52. Une petite progression est à noter pour ces deux catégorie d'outils et également pour les ressources linguistiques et la catégorie « Autres ».

Les deux autres phénomènes marquants sont la diminution du nombre d'outils de génération automatique et de gestion de terminologie.

4.2.2. Caractérisation des outils recensés

(a) Les outils selon la date de disponibilité et la commercialisation

On constate que les entreprises ont répondu sur des produits déjà commercialisés ou sur le point de l'être. Comme le montre le Tableau 7 ci-dessous, les outils qui ne sont pas disponibles actuellement appartiennent en général à des universités ou bien à des organismes publics. Pour le secteur industriel, seul l'outil de Synomia est concerné. Il ne sera disponible qu'à la mi-2002. En général, la disponibilité de certains outils réalisés dans des centres de recherche est prévue à plus long terme ou bien ces outils sont utilisés essentiellement en interne.

Tableau 7 : Disponibilité des outils

	Disponibles	**Non disponibles**	**TOTAL**
Industrie	174	1	175
Recherche	89	19	108
TOTAL	263	20	283

(b) Contraintes juridique, commerciale et technique des outils

La moitié des outils ne présente aucune contrainte, qu'elles soient d'ordre juridique, commerciale ou technique. Les outils proposés par les industriels sont déjà commercialisés ou sur le point de l'être. Parmi ceux qui présentent des contraintes, les deux principaux organismes concernés sont Xerox et IES-ADICOR. Pour le secteur de la recherche, les deux principaux organismes concernés sont le CRLT (Centre de Recherche en linguistique Lucien Tesnières) de l'université de Franche Comté et EDF.

Tableau 8 : *Contraintes juridique, commerciale et technique des outils*

Contraintes	Privé	Public	TOTAL
Pas de contraintes	105	36	141
Juridiques	1	16	17
Commerciales	1	1	2
Techniques	3	2	5
Juridiques et commerciales	4	4	8
Juridiques et techniques	1	2	3
Commerciales et techniques	0	1	1
Juridiques, commerciales et techniques	0	2	2
Info non disponible	60	44	104

(c) Les outils selon la maturité

Les applications permettant de traiter l'information électronique intègrent des produits/logiciels (élaborés et commercialisés) et des outils (ou modules), qui ont recours à des ressources linguistiques.

Chaque outil, produit/logiciel et application est à la base un prototype, c'est-à-dire le premier exemplaire d'un modèle (d'un mécanisme, etc.) construit, généralement à très peu d'unités exemplaires, à titre expérimental, et pour des démonstrations.

Tableau 9 : Catégorie des outils

	Prototype	Ressource	Outil	Application	Produit / Logiciel	Plusieurs catégories	Information non disponible
Industrie	0	5	23	9	87	8	33
Recherche	5	11	41	3	21	14	13
Total	5	16	64	12	108	22	46

(d) Les transferts de technologie

Les outils de l'industrie sont déjà commercialisés, et la plupart du temps, les organismes concepteurs en sont également les distributeurs. On relève également des outils du secteur public qui sont déjà commercialisés ou sur le point de l'être.

Tableau 10 : Transferts d'outils du secteur public vers le privé

Nom de l'outil	Nature de l'outil	Organisme fournisseur	Organisme distributeur
ALEx	12.2. Autres logiciels éducatifs	ENST de Bretagne	R.D.I
Sygmart	13. Outils d'aide à la construction d'outils ou de ressources linguistiques	LIRMM	IOSCA
KALI	8.5. Synthèse de la parole	CRISCO	Electrel
ILLICO	13. Outils d'aide à la construction d'outils ou de ressources linguistiques	LIF	Prologia-Groupe Air Liquide
KOMBE	12.2. Autres logiciels éducatifs	LIF	Prologia-Groupe Air Liquide
WinSnoori	8.7. Traitement du signal	LORIA	Babel Technologies
Camille	12.2. Autres logiciels éducatifs	LRL	Nathan CLE International
Ariane-G5	13. Outils d'aide à la construction d'outils ou de ressources linguistiques	CLIPS	Neurosoft
FIPS	2.2. Analyse syntaxique	LALT	LALT.ch
Dicovox	1.2.1.2. Bases de données lexicales bilingues ou multilingues	LALT	LALT.ch
FIPSTAG	2.3. Analyse morphologique	LALT	LALT.ch
FIPSVox	8.5. Synthèse de la parole (à partir du texte)	LALT	LALT.ch

Reacc	10.1.4. Système de réaccentuation automatique	RALI	Alis Technologies
SILC	8.4. Identification de la langue	RALI	Alis Technologies
Searchprocess Pro (Aurésys)	9.7. Recherche d'informations	CRRM	Intelligence Process
Pertinence (RAFI)	5. Outils de résumé automatique	Université de Nancy	Pertinence

Ceci étant, les informations sur les outils réalisés dans des centres de recherche, nous ont été transmises le plus souvent par l'organisme source, à savoir le laboratoire de recherche.

(e) Les outils selon les utilisateurs

Au total, 162 outils peuvent être utilisés par un utilisateur final, 108 par un intégrateur de logiciels et 85 par un chercheur. Un même outil peut avoir plusieurs types d'utilisateurs (colonnes 1 à 4). Les colonnes 5 à 7 montrent que certains outils ont des utilisateurs plus ciblés ; les outils uniquement destinés à la recherche proviennent du milieu même de la recherche alors que l'industrie favorise l'utilisateur final ou l'intégrateur de logiciels.

Tableau 11 : Utilisateurs potentiels

Utilisateurs[7] / **Secteur**	**(1) F, I, R**	**(2) F, I mais *R**	**(3) F, R mais *I**	**(4) I, R *F**	**(5) F seult**	**(6) I seult**	**(7) R seult**	**(8) Info non disponible**	**TOTAL**
Industrie	16	37	1	3	53	15	3	47	128
Recherche	17	5	17	14	16	1	14	24	84
TOTAL	33	42	18	17	69	16	17	71	212

[7] Notations : *F* pour Final, *I* pour Intégrateur et *R* pour la Recherche.

(f) La plate-forme standard

Tableau 12 : Machines

Machines / Secteur	(1) PC	(2) Mac	(3) Stat	(4) Gros syst	(5) PC, Mac	(6) PC, Mac, stat	(7) Tous	(8) PC, stat	(9) PC, gros syst	(10) Mac, stat	(11) PC, stat, gros syst	(12) Info non dispo	TOTAL
Industrie	92	0	2	1	13	4	4	12	1	0	0	46	175
Recherche	34	1	11	2	3	8	7	15	0	1	1	25	108
TOTAL	126	1	13	3	16	12	11	27	1	1	1	71	283

Tableau 13 : Systèmes d'exploitation

Système / Secteur	Unix	Ms-dos	Windows	Mac	Autre	Multi-systèmes	Info non dispo	TOTAL
Industrie	2	3	81	0	0	44	45	175
Recherche	17	1	27	0	9	29	25	108
TOTAL	19	4	108	0	9	73	70	283

Le type de matériel nécessaire et les logiciels requis pour les outils des secteurs public et privé semblent être assez standard. On dégage deux grandes tendances : un grand nombre d'outils requièrent un PC (64,6 %)[8] et fonctionnent principalement sous Windows (ainsi, 38 % des outils ne peuvent être utilisés que sous Windows)[9] ; un deuxième groupe d'outils fonctionnent plutôt sur des stations de travail sous Unix : il s'agit en majorité d'outils de centres de recherche. Les outils conçus uniquement pour des ordinateurs Macintosh sont assez rares. Nous n'en comptons qu'un seul, AMICAL. Précisons que les logiciels sont de

[8] Calcul effectué à partir des colonnes (1), (5), (6), (8), (9) et (11).

[9] Calcul effectué à partir de l'unique colonne « Windows ». Ce pourcentage augmente si l'on considère les outils disponibles sous Windows et pour d'autres systèmes d'exploitation (colonne « Multi-systèmes »).

manière générale disponibles pour diverses plates-formes : PC et/ou Macintosh et/ou station de travail.

(g) Les outils selon les langues

38,8 % des outils répertoriés sont multilingues (ils traitent donc trois langues ou plus), 32,2 monolingues et 12 % bilingues.

Tableau 12 : Monolingues, bilingues, multilingues

	Monolingue	**Bilingue**	**Multilingue 3 langues et +**	**Information non disponible**	**TOTAL**
Industrie	38	18	81	38	175
Recherche	56	16	29	7	108
TOTAL	94	34	110	45	283

uniquement du français. Il s'agit principalement d'analyseurs, d'outils de gestion de documents et de traitement de la parole.

Comme le montre le Tableau 13, les outils traitent principalement le français, l'anglais, l'espagnol, l'allemand, l'italien, le néerlandais et le portugais mais aussi des langues de l'Europe du nord. Viennent ensuite des langues appartenant à d'autres familles linguistiques comme le russe, le chinois, l'arabe, le coréen ou encore le japonais.

Tableau 13 : Langues traitées par les outils

Langues	**TOTAL**
Français	206
Anglais	105
Espagnol	66
Allemand	60
Italien	43

Langues	**TOTAL**
Japonais	8
Coréen	7
Islandais	6
Tchèque	6
Grec	6

Néerlandais	26	Irlandais	5
Portugais	25	Hongrois	5
Danois	15	Féroïen	4
Suédois	14	Estonien	3
Norvégien	13	Bulgare	3
Finnois	11	Basque	2
Russe	11	Croate	2
Chinois	10	Malais	2
Arabe	8	Turc	2
Polonais	8	Hébreu	2

Notons que les outils de BPI permettent de traiter cinq langues africaines à savoir le bambara, l'éwondo, le fulfulde, le lingala, le sango et le wolof.

Les langues citées dans la section « Autres » sont généralement des langues comme l'espagnol et l'anglais ou encore le portugais qui sont parlées dans des pays ou des régions particuliers. Par exemple, Speech cube traite l'espagnol castillan, l'anglais américain, l'espagnol sud-américain et le portugais brésilien.

Les outils qui traitent deux langues portent prioritairement sur le français et l'anglais.

Certains outils ont une couverture linguistique plus large et permettent par exemple de traiter plusieurs langues européennes. D'autres systèmes sont indépendants du choix de la langue comme par exemple WORLDTREK et WinSnoori. Certains outils définissent les langues traitées par des caractéristiques typographiques ou linguistiques ; par exemple le système d'acquisition de terminologie ANA, de l'IRIN, traite les langues non agglutinantes, alors que les outils du LILLA (ZSTATION, ZTEXT et ZTERMINO), Solare et Multitrans ne traitent les langues à alphabet latin. Ces systèmes multilingues appartiennent à différents domaines. On y relève essentiellement des outils de gestion électronique de documents et de traitement de la parole alors que les systèmes de traduction sont généralement bilingues mais existent pour différentes paires de langues.

5. Exploitation des données

Une des tâches importantes de ce projet était consacrée à la réalisation d'un site Web particulier permettant d'accéder à une liste d'acteurs appartenant au domaine de l'ingénierie linguistique et d'outils de traitement automatique de la langue mais aussi permettant une collecte d'informations concernant les organismes et décrivant les outils recensés de façon conviviale et attrayante mais surtout de façon interactive. Ce site devrait permettre d'avoir un répertoire toujours à jour.

5.1. Consultation des données

L'un des objectifs de ce projet, comme nous venons de le souligner, était de pouvoir donner accès à des informations concernant des organismes du domaine de l'ingénierie linguistique et les outils qu'ils développent et/ou distribuent.

Les informations qui seront disponibles sur le site Web seront bien évidemment les informations obtenues dans le cadre des deux enquêtes mais elles seront synthétisées. En d'autres termes, nous effectuerons une sélection des informations.

Plusieurs inventaires d'outils ou / et de ressources sont aujourd'hui accessibles, notamment sur le Web. Certains organismes tels ELDA (Agence pour l'évaluation et la distribution de ressources linguistiques) et le LDC (Linguistic Data Consortium) proposent ainsi des inventaires de ressources linguistiques. D'autres proposent des inventaires d'outils de TAL pour lesquels plusieurs types d'informations sont disponibles. Citons entres autres, les travaux de la DGLF (Délégation Générale à la Langue Française)[10], du DFKI [11], de l'OWIL (Office Wallon des Industries de la Langue)[12], du CEVEIL et du RIOFIL[13] et de l' Atala[14]. Ces sites offrent la possibilité d'accéder aux informations de différentes manières soit en sélectionnant une catégorie d'outils (traduction automatique, analyse linguistique, etc.) soit en sélectionnant

[10] À la découverte de l'ingénierie linguistique en France (http://www.culture.gouv.fr/culture/dglf/rifal/garde.htm).

[11] Natural Language Software Registry (http://registry.dfki.de/).

[12] Répertoire de l'inforoute et du Traitement Informatique des Langues en Région Wallonne et Bruxelles Capitale (http://www.owil.org/fr_repertoire.htm).

[13] VOIL ! Vitrine des Outils Inforoutes et Langues de la Francophonie (http://199.84.130.137/).

[14] Répertoire d'outils pour le Traitement Automatique des Langues (http://www.biomath.jussieu.fr/ATALA/outil/).

un organisme. Un classement alphabétique des outils est également proposé sur la plupart de ces sites, de même qu'une interface de recherche.

Plusieurs solutions sont ainsi possibles :

- une première solution pourrait être de présenter les données sous la forme de deux listes : une liste d'outils classés par ordre alphabétique et une liste d'organismes classés également par ordre alphabétique. Des liens hypertextes permettraient ensuite d'accéder aux descriptions des outils et des organismes ;
- une seconde possibilité pourrait être de classer les outils selon leur type, et des liens hypertextes permettraient d'accéder aux descriptions ;
- une troisième solution serait de laisser la possibilité à l'utilisateur d'effectuer des requêtes selon un certain nombre de critères prédéfinis. Les recherches pourraient par exemple porter sur des critères simples (catégorie de l'outil, langue(s) traitée(s), disponibilité, plate-forme, etc.) ou une combinaison de critères ;
- une dernière solution dans la continuité de la précédente serait de permettre une interrogation en langage naturel. L'intégration d'un moteur de recherche existant et éventuellement d'un outil de traduction automatique permettraient d'exploiter pleinement les données. Il serait même envisageable d'élargir le champs d'investigation de cette étude afin de couvrir tout la domaine de l'ingénierie linguistique en France de façon plus générale : organismes, membres, activités, outils, projets de recherches en cours, colloques, publications, collaborations internationales, etc.

La troisième solution semble réalisable et tout à fait appropriée, la meilleure solution envisageable à plus long terme, mais plus complexe, serait bien sûr la dernière.

Cet inventaire sera également présenté sous la forme d'un catalogue sur CD-Rom, qui sera diffusé auprès de la communauté de l'ingénierie linguistique.

5.2. Saisie des données

L'objectif est de donner la possibilité aux acteurs du domaine de l'ingénierie linguistique de saisir des données les concernant ou concernant les outils qu'ils développent dans des formulaires simples, structurés et attrayants. C'est ce que propose l'ensemble des sites cités dans le paragraphe précédent à l'exception du site de la DGLF.

Nous avons utilisé le même formulaire que celui qui a été utilisé dans le cadre de l'enquête. Il a néanmoins été un peu allégé afin de faciliter la saisie des données. Une fois les informations saisies, elles sont envoyées à un contact ELDA qui les valide ou ne les valide pas, l'objectif étant d'éviter tout discours trop commercial.

6. Conclusion

Le projet consistait à mettre à jour l'inventaire des outils de traitement automatique de la langue et des ressources linguistiques correspondant à l'offre française, réalisé en 1999. 257 outils et 26 ressources ont été recensés.

Ce travail peut servir de support aux recherches et aux travaux de chercheurs et aux industriels, désirant recourir aux technologies linguistiques. Il peut également susciter des collaborations entre centres de recherches et sociétés de développement de technologies. Cette description des outils peut, par ailleurs, constituer un élément essentiel dans leur éventuelle intégration dans des applications, telles que, par exemple, la sécurité de l'information ou la diffusion de l'information.

Sa mise en ligne permet de collaborer à cette tâche de promotion de l'ingénierie linguistique tandis que le formulaire de saisie permet de garder les données continuellement à jour. Il est en effet envisagé de réaliser une mise à jour régulière.

7. Annexes

Annexe 1 : Typologie des outils de TAL

Typologie des outils de TAL

Cochez la case correspondante ou reporter l'information dans le champ *Type* de la rubrique (2) *IDENTIFICATION DE L'OUTIL OU DE LA RESSOURCE LINGUISTIQUE*.

1. RESSOURCES LINGUISTIQUES

- **1.1. Ressources orales** ☐
  - 1.1.1. Bases de données acoustiques (téléphonie) ☐
  - 1.1.2. Bases de données acoustiques (microphone) ☐
  - 1.1.3. Lexiques de prononciations ☐
- **1.2. Ressources écrites** ☐
  - 1.2.1. Bases de données lexicales ☐

1.2.1.1. Bases de données lexicales monolingues ☐
1.2.1.2. Bases de données lexicales bilingues ou multilingues ☐
1.2.2. Bases de données terminologiques ☐
1.2.2.1. Bases de données terminologiques monolingues ☐
1.2.2.2. Bases de données terminologiques bilingues ou multilingues ☐
1.2.3. Corpus textuels ☐
1.2.3.1. Corpus textuels monolingues ☐
1.2.3.2. Corpus textuels bilingues ou multilingues ☐

1.3. Ressources multimodales/multimédia ☐

1.4. Grammaires ☐

1.5. Autres ☐

2. OUTILS D'ANALYSE LINGUISTIQUE

2.1. Analyse sémantique et/ou pragmatique ☐
2.2. Analyse syntaxique ☐
2.3. Analyse phonétique ☐
2.4. Analyse morphologique ☐
2.4.1. Analyse flexionnelle (ou lemmatiseur) ☐
2.4.2. Analyse dérivationnelle ☐

3. OUTILS DE GENERATION AUTOMATIQUE

3.1. Génération de formes fléchies et/ou dérivées ☐
3.2. Génération de phrases ☐
3.3. Génération de textes ☐
3.4. Génération de documents multimédia ☐

4. OUTILS DE TRADUCTION AUTOMATIQUE ET ASSISTEE

4.1. Traduction automatique ☐
4.2. Traduction assistée par ordinateur ☐
4.3. Aide à la traduction (mémoires de traduction, dictionnaires interactifs) ☐
4.4. Post-édition automatique ☐
4.5. Traduction vocale ☐
4.6. Aligneur de corpus ☐
4.7. Localisation assistée ☐

5. OUTILS DE RESUME AUTOMATIQUE ☐

6. OUTILS DE COMPREHENSION DU LANGAGE NATUREL ☐

7. OUTILS DE GESTION DE TERMINOLOGIE

7.1. Extraction terminologique ☐
7.2. Consolidation terminologique ☐
7.3. Autres ☐

8. OUTILS DE TRAITEMENT AUTOMATIQUE DE LA PAROLE

- 8.1. Reconnaissance automatique de la parole ☐
- 8.2. Dictée vocale ☐
- 8.3. Identification et vérification du locuteur ☐
- 8.4. Identification de la langue ☐
- 8.5. Synthèse de la parole (à partir du texte) ☐
- 8.6. Dialogue Homme-Machine ☐
- 8.7. Analyse et manipulation du signal ☐
- 8.8. Analyse prosodique ☐
- 8.9. Autres applications de technologies vocales ☐

9. OUTILS DE GESTION ELECTRONIQUE DE DOCUMENTS ET RECHERCHE D'INFORMATIONS

- 9.1. Segmentation de textes (en paragraphes, phrases, mots, caractères) ☐
- 9.2. Statistiques lexicales ☐
- 9.3. Gestion et reconnaissance de format électronique de documents ☐
- 9.4. Classement ou classification de documents ☐
- 9.5. Identification de la langue ☐
- 9.6. Indexation de documents ☐
- 9.7. Recherche d'informations ☐
- 9.8. Aide à la recherche d'informations ☐
- 9.9. Interrogation en langage naturel (LN) ☐
- 9.10. Filtrage d'informations ☐
- 9.11. Routage de documents ☐
- 9.12. Extraction d'informations ☐
- 9.13. Système de navigation documentaire ☐

10. OUTILS D'AIDE A LA REDACTION

- 10.1. Outils de détection et de correction automatique d'erreurs ☐
  - 10.1.1. Correcteur orthographique / lexical ☐
  - 10.1.2. Correcteur grammatical ☐
  - 10.1.3. Correcteur stylistique ☐
  - 10.1.4. Système de réaccentuation automatique ☐
  - 10.1.5. Aide orthographique ☐
- 10.2. Prédicteur lexical ☐
- 10.3. Traitement de texte avancé ☐
- 10.4. Langages contrôlés ☐
- 10.5. Autres ☐

11. OUTILS DE TRAITEMENT OPTIQUE DE CARACTERES

- 11.1. Reconnaissance de caractères imprimés ☐
- 11.2. Reconnaissance de caractères manuscrits ☐
- 11.3. Autres ☐

12. OUTILS D'ENSEIGNEMENT INTELLIGEMMENT ASSISTE PAR ORDINATEUR (EIAO)

12.1. Outils d'apprentissage des langues ☐

12.2. Autres logiciels éducatifs ☐

13. OUTILS D'AIDE A LA CONSTRUCTION D'OUTILS OU DE RESSOURCES LINGUISTIQUES ☐

(générateur d'application, interfaces de visualisation, aide à l'acquisition de ressources orales)

Annexe 2 : Questionnaire

Questionnaire

Enquête sur les outils et ressources linguistiques
pour le traitement automatique des langues en France

Apportez des modifications si nécessaires

1. VOS COORDONNEES

Organisme			
Nom			
Raison sociale			
Type	☐ Entreprise ☐ Université ☐ Organisme public		
Adresse	Rue :		
	Code postal :	Ville :	Pays :
Site Web			

Contact		
	Personne contactée	Personne à contacter (*si différente*)
Nom		
Courrier électronique		
Téléphone		
Télécopie		

Autre(s) organisme(s) ayant participé à la réalisation de l'outil ou de la ressource linguistique			
Nom			
Raison sociale			
Contact			
Type	☐ Entreprise ☐ Université ☐ Organisme public		
Adresse	Rue :		
	Code postal :	Ville :	Pays :
Site Web			

2. IDENTIFICATION DE L'OUTIL OU DE LA RESSOURCE LINGUISTIQUE

Nom	
Version	
Libellé complet	

Catégorie	☐ Ressource linguistique ☐ Prototype ☐ Outil ☐ Produit / logiciel ☐ Application
Type *(voir liste en annexe)*	
Langue(s) traitée(s)	☐ Français ☐ Autres, *précisez* :
Disponibilité	☐ Est disponible ☐ Sera disponible à partir de :
Utilisateurs potentiels	☐ Utilisateur final ☐ Intégrateur de logiciels ☐ Chercheur

3. ASPECTS TECHNIQUES

Support	
Type	☐ Disquette(s) ☐ CD-Rom ☐ Téléchargement ☐ DVD
	☐ Autres, *précisez* :
Taille	….Mo

Système requis	
Type	☐ Pc, *précisez le* Processeur : ☐ Macintosh, *précisez le* Processeur : ☐ Station de travail, *précisez* : ☐ Gros Système, *précisez* :
Mémoire vive	Minimum :…Mo Recommandée :…Mo
Espace nécessaire sur le disque	….Mo
Système d'exploitation	☐ MS-DOS ☐ Windows, *précisez* : ☐ Apple OS, *précisez* : ☐ Unix, *précisez* : ☐ Autres, *précisez* :

Intégration de l'outil ou de la ressource linguistique
☐ Autonome
☐ Intégré(e) dans l'application suivante :
☐ Intégrable dans les applications suivantes :

Documentation
☐ Disponible en français
☐ Disponible dans d'autres langues, *précisez* :

Existe-il des contraintes pour la commercialisation de votre outil ou ressource linguistique?			
☐ Non			
☐ Oui, d'ordre :	☐ juridique	☐ technique	☐ commercial
☐ Déjà commercialisé(e), depuis le :			

Coordonnées du distributeur (si différentes de l'organisme) :	
Nom	
Raison sociale	
Contact	
Type	☐ Entreprise ☐ Université ☐ Organisme public

Adresse	Rue :		
	Code postal :	Ville :	Pays :
Site Web			

4. DESCRIPTION DETAILLEE DE L'OUTIL OU DE LA RESSOURCE LINGUISTIQUE

[Ancienne description]

5. COMMENTAIRES

[Anciens commentaires]