ELDA

European **L**anguage Resources **D**istribution **A**gency

Agence **E**uropéenne de **D**istribution des **R**essources **L**inguistiques

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

RESUME DE L'ETUDE

Sommaire

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 des Langues (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.

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

Pour mener à bien l'enquête de 1999, une typologie des outils de TAL et un formulaire de description couvrant plusieurs critères (nom de l'outil, type de l'outil, plates-formes matérielle et logicielle requises, contraintes de commercialisation, etc.) ont été créés et une liste de prospects, c'est-à-dire d'organismes et/ou de personnes à contacter, constituée. Pour la seconde enquête, ces documents n'ont fait l'objet que de légères modifications.

La stratégie employée lors de la seconde enquête était la même que celle définie dans la première. Il s'agissait d'adresser individuellement aux organismes identifiés une lettre expliquant notre démarche et un formulaire de description et de procéder à des relances régulières lorsqu'ils ne répondaient pas. Lorsqu'ils étaient destinés à de nouveaux organismes, les formulaires étaient pré-remplis dans leur partie « Vos coordonnées ». Lorsqu'ils étaient destinés aux organismes déjà contactés, les formulaires comprenaient les informations recueillies lors de la première enquête.

3. Les données

3.1. Les réponses reçues et les données recueillies par ELDA

Lors de la première enquête, sur 253 organismes prospectés, seuls 83 d'entre eux 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. Nous avons répertorié et décrit des outils que les organismes prospectés n'avaient pas signalés. Nous avons également mené des recherches sur des outils pour lesquels nous disposions d'informations insuffisantes.

3.2. Les outils recensés

Cette partie se propose de faire un bilan concernant les données obtenues. Elle fait rapidement le point sur les caractéristiques suivantes : type d'outil, date de disponibilité et de commercialisation, maturité, disponibilité juridique, commerciale et technique, transferts de technologie, types d'utilisateurs, plate-forme standard et langues traitées.

3.2.1. Typologie des outils recensés

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.

3.2.2. 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. 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.

(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. Lorsqu'il y a des contraintes, elles sont essentiellement juridiques dans le secteur public (c'est le cas pour 16 outils, sur les 64 outils présentant des contraintes), plutôt juridiques et commerciales pour le secteur privé. 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.

(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.

Nous avons compté 108 produits/logiciels, 64 outils, 16 ressources linguistiques, 12 applications et 5 prototypes. 22 items appartiennent à deux ou plusieurs catégories. Les produits/logiciels et les applications sont principalement issus du secteur privé. Par contre, les ressources linguistiques, les outils et les prototypes appartiennent au secteur public.

(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 : ALEx (ENST de Bretagne), distribué par R.D.I, KALI (CRSICO), distribué par Electrel, etc.

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. Certains outils ont des utilisateurs plus ciblés ; les outils uniquement destinés à la recherche proviennent du milieu de la recherche alors que l'industrie favorise l'utilisateur final ou l'intégrateur de logiciels.

(f) La plate-forme standard

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 %) et fonctionnent principalement sous Windows (ainsi, 38 % des outils ne peuvent être utilisés que sous Windows) ; 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 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. 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.

Notons que les outils de dépouillement automatique de corpus de textes de terminologie de Progiciels Bourbeau Pinard permettent de traiter cinq langues africaines à savoir le bambara, l'éwondo, le fulfulde, le lingala, le sango et le wolof.

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. Certains outils définissent les langues traitées par des caractéristiques typographiques ou linguistiques.

4. 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.

4.1. Consultation des données

L'un des objectifs de ce projet 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 d'ELDA seront les informations obtenues dans le cadre des deux enquêtes, synthétisées.

Plusieurs solutions étaient 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 les organismes selon le domaine en technologies de la langue dans lequel ils s'inscrivent, 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.

Les deux premières solutions, combinées, semblaient les plus facilement réalisables et tout à fait appropriées. La meilleure solution, envisagée à plus long terme, et plus complexe, sera de combiner les deux dernières solutions.

4.2. 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.

Nous envisageons d'utiliser une version allégée du formulaire envoyé aux organismes dans le cadre de l'enquête. Une fois les informations saisies, elles seraient envoyées à un contact ELDA qui les validerait ou ne les validerait pas, l'objectif étant d'éviter tout discours trop commercial.

5. 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.

Cette étude a plusieurs vocations : elle peut servir de support aux recherches et aux travaux de chercheurs et aux industriels, désirant recourir aux technologies linguistiques. Elle 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 de traitement de l'information.