La place de l'arabe dans le programme Technolangue

Stéphane Chaudiron
Université de Lille 3
Ministère délégué à la Recherche

Résumé :

Technolangue est le programme national français de soutien au domaine des technologies de la langue lancé en 2002 par les ministères chargés de la Recherche, de l'Industrie et de la Culture. Il a concerné quatre volets : la création de ressources linguistiques et logicielles de base, l'organisation de huit campagnes d'évaluation, la participation aux instances internationales de normalisation et la création d'un portail Web destiné à diffuser les informations concernant le domaine des technologies de la langue. Cet article présente les principaux résultats du programme, notamment les projets qui se sont intéressés au traitement de la langue arabe.

1. Introduction

À la suite d'un rapport remis au Premier Ministre en 2000 mettant en évidence le rôle majeur des technologies de la langue (TL) dans la société de l'information, le programme *Technolangue* a été lancé en

avril 2002 comme programme national de soutien aux technologies de la langue, tant écrite qu'orale.

Technolangue a été pensé en étroite complémentarité des programmes de R&D concernant les secteurs des technologies de l'information et de la communication, en particulier les Réseaux de recherche et d'innovation technologique (RRIT) dans le secteur des télécommunications (RNRT), du logiciel (RNTL) et de l'audiovisuel et du multimédia (RIAM). En complément des RRIT chargés de financer les projets de R&D, *Technolangue* a été une action transversale destinée à créer des ressources linguistiques et logicielles de base, à assurer leur diffusion de manière pérenne, à encourager la participation aux instances de normalisation, à organiser des campagnes d'évaluation et à mettre en place un dispositif de veille spécialisé dans les TL.

Dans le cadre de *Technolangue*, plusieurs langues partenaires du français ont été impliquées dans différents projets comme l'italien, l'espagnol, l'allemand, l'anglais, le grec et d'autres encore. Plusieurs projets de création de ressources linguistiques ou d'évaluation de logiciels ont également concerné la langue arabe.

Cet article présente les principaux résultats issus des quatre volets du programme et insiste particulièrement sur les projets qui ont concerné le traitement de l'arabe.

2. Le contexte du programme

Le programme *Technolangue* a été lancé en 2002 et les derniers résultats ont été présentés en 2006. Un comité de pilotage de 15 membres représentant à parité les domaines de la parole et de l'écrit et les secteurs

industriel et académique a été chargé de définir les axes, superviser l'évaluation des propositions reçues à l'issue de l'appel à projets et s'assurer du bon déroulement des projets.

Le budget total consacré à ce programme a été de 7,5 millions d'euros provenant des trois ministères en charge de la Recherche, de l'Industrie et de la Culture (via le CNC pour ce dernier). L'effort global d'investissement, fourni à la fois par les ministères et le secteur industriel, atteint 11,5 M€, dans la mesure où de nombreux projets ont été financés sur la base d'un coût partagé.

Quatre volets ont concerné *Technolangue* :

- Le premier a visé à encourager la production et la diffusion de ressources linguistiques et d'outils logiciels de base pour le traitement de la langue. Ce volet avait pour objectif la création d'une « boite à outils » contenant les ressources linguistiques et logicielles minimales pour le traitement automatisé du français. Un aspect central a concerné la diffusion la plus large possible des ressources créées.
- Le deuxième volet avait pour but de financer l'organisation de campagnes d'évaluation des technologies de traitement de la langue orale et écrite.
- Le troisième volet a visé à encourager la participation française aux instances internationales de standardisation et normalisation et assurer la diffusion des informations auprès de la communauté scientifique et industrielle.
- Enfin, le dernier volet a concerné la création d'un portail de veille sur Internet destiné à diffuser toutes les

informations, scientifiques, techniques, industrielles, concernant le domaine des TL.

Fin 2002, 27 projets sur 52 ont été retenus à l'issue de la phase d'évaluation et 21 projets ont été finalement financés. Plus de 90 participants différents ont été impliqués dans les projets, 33 provenant de l'industrie, 39 du monde académique, 11 divers (associations de loi 1901, DGA, CEA…) et 11 participants étrangers (Bell Labs (USA), NII (Japon), EPFL et LATL (Suisse), RALI (Canada)…), qui ont participé au programme avec leur propre financement.

3. Résultats du programme

Les 21 projets financés ont ensuite été regroupés en 5 « clusters » en fonction de leurs thématiques, des statuts des partenaires et des sources de financement. La section ci-dessous présente un bref descriptif des résultats obtenus par chacun des projets constituant les « clusters » ; une présentation plus détaillée des résultats est disponible sur le portail *Technolangue* (www.technolangue.net) et sur les sites des projets.

3.1. Création de ressources linguistiques et logicielles de base

Les trois premiers « clusters », AGILE, NEOLOGOS et DICTIONNAIRE, ainsi que le projet CARMEL ont visé la création et la diffusion de ressources de base. Au sein de AGILE, quatre projets ont été financés :

3.1.1. TILT

Coordonnateur du projet: ATILF et AFNOR

Le principal résultat du projet a été la création d'un corpus balisé XML d'environ 1 000 normes alignées en français-anglais provenant de l'AFNOR qui a ensuite aidé au développement d'applications linguistiques de haut niveau telles des outils de traduction, des outils de recherche d'information pour les documentalistes, les utilisateurs des normes et les scientifiques. Le projet a également permis de produire :

- Une méthode de balisage pouvant être appliquée à d'autres textes de normes,
- Une liste bilingue de termes qui peuvent être utilisés pour enrichir des dictionnaires techniques et un ensemble de groupes syntaxiques qui peut également enrichir les mémoires des logiciels de traduction,
- Des outils de recherche utilisant les balises et permettant la recherche de termes en contexte, e nombre d'occurrences, les exceptions dans les normes…,
- Des prototypes tels que par exemple un moteur de recherche sémantique et un résumeur automatique.

Site:http ://stella.atilf.fr/dendien/recherche- nor.htm

3.1.2. ALIZÉ

Coordonnateur du projet: consortium ELISA

ALIZÉ est une « boîte à outils » pour la reconnaissance du locuteur, gratuite et ouverte. Elle a été développée par le consortium ELISA en fonction de trois objectifs : rester simple et pédagogique afin de pouvoir être utilisée par les étudiants pour leur formation ou la recherche, avoir un niveau de performance correspondant à l'état de l'art en terme de taux d'erreur comme en

utilisation de ressources, et aider au développement de démonstrateurs ou d'applications industrielles.
Site:http ://www.lia.univ-avignon.fr/heberges/ALIZE/

3.1.3. OURAL

Coordonnateur du projet: Sinequa

Le projet a donné trois types de résultats : des lexiques, des corpus et des outils logiciels, l'ensemble étant diffusé sous licence GNU. Différents types de lexiques ont été créés tels qu'une base lexicale du français d'environ 160 000 formes (44 000 lemmes, indication des fréquences d'apparition à l'oral et à l'écrit), des lexiques d'anagrammes, de prénoms, d'homographes, etc. Différents corpus ont également été créés : un corpus écrit au format TEI (environ 10 000 mots) et quatre corpus constitués de transcriptions d'entretiens et de dialogues Enfin, plusieurs outils ont été développés et sont diffusés : segmenteur, étiqueteur, extracteur de concept, outil de filtrage, résumeur…

3.1.4. WATSON

Coordonnateur du projet: Lingway

Le projet Watson a développé, adapté, intégré et/ou généralisé des outils pour permettre la structuration logique de pages web, la reconnaissance des entités nommées, le balisage de textes ; des étiqueteurs, extracteurs, outils de catégorisation, de résolution de co-référence et de résumé automatique ont aussi été créés. Ces modules logiciels ont été intégrés dans les solutions applicatives de Lingway, en particulier la solution Lingway KM, mais peuvent également être utilisées de manière autonome. Ils ont été conçus pour être robustes

et offrir une performance maximale sur de gros volumes de données. Les résultats sont disponibles à des fins pédagogiques et de recherche et leur diffusion est directement assurée par Lingway.

3.1.5. NEOLOGOS

Coordonnateur du projet: TELISMA

Dans le cadre de ce projet, deux bases de données parole nouvelles ont été développées (**Paidialogos** et **Idiologos**) destinées à être utilisées pour le traitement automatique de la parole.

PAIDIALOGOS est une base de données de voix d'enfants et d'adolescents livrée au format SpeechDat qui un format traditionnellement utilisé dans la distribution de bases de données vocales. La base comprend plus de 37000 enregistrements téléphoniques produits par 1 010 enfants, respectant des contraintes de répartition par genre, âge et région. Du point de vue de la couverture phonétique, la base couvre l'ensemble des phonèmes de la langue française en respectant également la fréquence d'usage.

La base IDIOLOGOS comprend une première base, dite « bootstrap » conçue dans le but de constituer un corpus utilisé pour la sélection de locuteurs de référence et permettre d'effectuer des apprentissages de reconnaissance de la parole pour les adultes de langue maternelle française. Elle est le résultat de l'enregistrement au format *SpeechDat* par le réseau téléphonique fixe de 1 000 locuteurs ayant enregistré chacun les mêmes 45 phrases phonétiquement riches et 5 items différents d'un locuteur à l'autre. La seconde partie

de la base comprend les 500 enregistrements des 200 locuteurs sélectionnés, soit près de 100 000 enregistrements.

Dans le cadre du « cluster » **DICTIONNAIRE**, quatre projets ont été financés :

3.1.6. EURADIC

Coordonnateur du projet: CEA et CNRS

L'objectif du projet **EurADiC** a été double. Il s'agissait d'abord de créer ou compléter plusieurs dictionnaires monolingues et bilingues pour la langue générale ou de spécialité. Des dictionnaires monolingues de langue générale ont été développés pour le français, l'allemand, l'anglais, l'espagnol et l'italien ; le volume des ressources oscille entre 45 000 lemmes ou parties de discours pour le dictionnaire italien et plus de 155 000 entrées pour le dictionnaire anglais. Des dictionnaires bilingues de langue générale ont également été créés pour un minimum de 90 000 liens bilingues pour les langues suivantes : français-allemand, français-anglais, français-arabe, français-espagnol et français-italien. Un dictionnaire terminologique dans le domaine du sport a enfin été développé pour le français, l'anglais, l'allemand, l'espagnol, le grec et l'arabe.

Le deuxième objectif du projet a été de créé un dictionnaire monolingue arabe et un corpus textuel bilingue français-arabe. La taille du dictionnaire, entièrement voyellé et balisé, est de 105 000 mots. Le corpus comprend 42 paires de textes alignés au niveau de la phrase. Les deux résultats proviennent du journal *Le Monde Diplomatique*.

3.1.7. ATONANT

Coordonnateur du projet: EADS

Le projet a permis la création d'une « boîte à outils » destinée à faciliter le développement de ressources sémantiques, principalement des ontologies. La boîte à outils est constituée de cinq outils indépendants ; le processus global est d'aspirer le web à partir d'URL prédéfinis de façon à constituer un corpus, de normaliser celui-ci au format TEI et d'aider à la création d'une ontologie à partir du corpus. Chaque outil correspond à une étape du processus qui permet de partir des données brutes (ici les pages Web) pour arriver à un corpus sémantiquement enrichi. La boîte à outils a été développée et évaluée dans le cadre d'une application opérationnelle d'enrichissement de ressources dans le domaine médical.

Site: http://atonant.insa-rouen.fr

3.1.8. Noms Propres

Coordonnateur du projet: Université de Tours

L'objectif principal du projet a été de créer des ressources multilingues pour la traduction des noms propres. Ces ressources n'ont pas été conçues à partir de dictionnaires bilingues ou multilingues mais à partir de dictionnaires monolingues partageant les mêmes concepts et avec des liens interlingues. Pour cela, une ontologie a été développée, structurée en deux parties : une partie monolingue et une partie multilingue. Le noyau de la partie multilingue est le concept qui joue le rôle d'identifiant interlingue. La base de données *Prolexbase*, qui est accessible en ligne, dépasse les 53 000 noms propres (environ 122 500 formes fléchies)

pour le français et comporte plus de 400 relations de synonymie, plus de 2 200 relations d'accessibilité, plus de 44 000 liens de méronymie. Environ 700 noms propres sont traduits en anglais, italien, allemand, espagnol, flamand, portugais et serbe.
Site: http://tln.li.univ-tours.fr/tln_prolex/prolex.php

3.1.9. LEXITEC

Coordonnateur du projet: Softissimo

Le projet **Lexitec** a visé à la création de dictionnaires bilingues de spécialité au moyen d'outils d'acquisition terminologique, dans des domaines où manquent des ressources lexicales structurées et facilement utilisables. Le projet a été conduit dans le cadre d'un partenariat entre terminologues, spécialistes de la normalisation et prescripteurs de technologies. Les dictionnaires obtenus sont disponibles dans des formats d'échange standardisés et concernent les domaines suivants : l'aéronautique (6 923 entrées pour le sens FR⇒EN et 4 271 pour le sens EN⇒FR), l'automobile (3 382 pour le sens FR⇒SP et 2 238 le sens SP⇒FR), les affaires (5 814 entrées pour FR⇒SP et 2 016 pour SP⇒FR), des expressions idiomatiques (4 264 entrées pour FR⇒GE, 3 766 pour GE⇒FR, 1 334 entrées pour FR⇒EN et 1 064 entrées pour EN⇒FR) et la mécanique (4 554 entrées pour FR⇒EN, 3 345 entrées pour EN⇒FR).

3.1.10 CARMEL

Coordonnateur du projet: Université d'Avignon.

Le projet CARMEL a permis la constitution d'un corpus multilingue aligné couplé à un ensemble d'outils

d'exploration. Le corpus est constitué d'une collection d'œuvres littéraires du XIXe siècle (récits de voyages) dans quatre langues européennes : le français, l'anglais, l'espagnol et l'italien. Les textes, mis au format XML, ont été enrichis par l'ajout d'annotations morphosyntaxiques (étiquetage et lemmatisation), sémantiques et thématiques pour en faciliter l'exploitation. Un site permet d'explorer le corpus à travers une interface basée sur le moteur de recherche *Intuition* développé par *Sinequa*. Ce site permet également de télécharger gratuitement les textes libres de droit (la plupart des textes originaux certaines traductions), ainsi que certains outils pour l'annotation et l'alignement. Les outils accompagnant le corpus permettent l'alignement de nouveaux textes ainsi que la visualisation de bi-textes et de leurs annotations.
Site : http://www.projetcarmel.org

3.2. Des campagnes d'évaluation

Dans le cadre du « cluster » EVALDA, huit campagnes d'évaluation des technologies de la langue ont été organisées.

3.2.1. ARCADE 2

La campagne **ARCADE 2** visait à explorer les techniques d'alignement de textes multilingues à travers une évaluation fine des techniques existantes et le développement de nouvelles méthodes d'alignement. La campagne a permis d'évaluer les outils d'alignement au niveau de la phrase et au niveau du mot. La tâche d'évaluation portant sur l'alignement au niveau de la phrase s'est déroulée selon le scénario suivant : un

ensemble de textes parallèles segmentés en phrases étaient proposés aux participants à l'évaluation qui devaient ensuite fournir comme résultat l'alignement des phrases. Pour l'alignement de phrases, deux tâches ont été proposées aux participants : l'alignement de textes en 5 langues européennes et l'alignement de textes rédigés en 6 langue extra-européennes.

La tâche d'alignement au niveau des mots a comporté une sous-tâche particulière consistant à l'identification d'entités nommées traduites dans des textes parallèles français-arabe.

Pour les langues européennes, un corpus issu des *Journaux Officiels de la Communauté* européenne (JOC) a été utilisé (environ 5 millions de mots également répartis en anglais, français, allemand, italien et espagnol). Le même sous-ensemble en anglais, allemand, italien et espagnol était aligné sur l'équivalent français au niveau de la phrase et du paragraphe.

Pour les langues non européennes, un corpus d'articles issu du mensuel *Le Monde Diplomatique* a été utilisé. Ce corpus contenait 150 textes en arabe alignés avec le français au niveau de la phrase, 50 paires de textes en russe, chinois, japonais, grec et persan alignés sur le français qui jouait le rôle de langue pivot. Un sous-ensemble de textes français contenant des entités nommées étiquetées a été utilisé pour la tâche d'alignement au niveau des mots.

3.2.2. CESART

La campagne **CESART** a concerné l'évaluation de systèmes d'acquisition de ressources terminologiques. Le projet a permis de concevoir et valider un protocole

d'évaluation destiné à comparer différents systèmes d'extraction utilisés pour la création de ressources terminologiques ou l'extraction de relation sémantiques. Le projet a aussi abouti à la création de ressources linguistiques de haute qualité telles que des corpus spécialisés et des outils de mesure de performance.

Pour la tâche d'extraction de termes, un corpus de textes a été fourni aux participants qui devaient retourner une liste de candidats termes classés par ordre de pertinence, avec des indications concernant leur fréquence d'apparition dans le corpus, les variantes et le contexte d'apparition. Un processus automatique a dabord été utilisé pour comparer les sorties des systèmes avec une liste de référence générée à partir d'un thésaurus du domaine existant. Tous les termes identiques ont été considérés comme pertinents et les autres ont été soumis au jugement d'un expert.

Pour la tâche d'extraction sémantique, les participants ont reçu le même corpus déjà utilisé pour la tâche d'extraction de termes ainsi que les résultats fournis par les systèmes à l'issue de cette première tâche. Il était demandé aux participants de cette seconde tâche de fournir une liste établissant les relations de synonymie entre les candidats termes. Pour cette tâche, on a mesuré la correspondance entre les résultats fournis par les systèmes et les liens de synonymie indiqués dans un thesaurus existant.

Deux corpus de spécialité ont été créés spécifiquement pour la campagne d'évaluation. Le premier est un corpus d'environ 9 millions de mots correspondant à des documents en français issus du site web de Santé Canada (http://www.hc-sc.gc.ca). Le

second, d'environ 500 000 mots, est constitué de textes de la revue scientifique *Spirale*, consacrée aux sciences de l'éducation. Un troisième corpus, correspondant au sous-ensemble français extrait des Questions et Réponses du *Journal Officiel de la Communauté Européenne* (JOC) a également été utilisé pour masquer les deux précédents corpus mais les résultats correspondant à ce corpus n'ont pas été évalués.

3.2.3 CESTA

Le projet **CESTA** a consisté en deux campagnes d'évaluation portant sur la traduction automatique (TA). Les langues concernées ont été l'anglais et l'arabe en tant que langues source et le français en tant que langue cible.

Le projet a d'abord discuté les différentes métriques classiquement utilisées pour l'évaluation en TA, telles que les métriques *BLEU* du NIST, *mWER* et *mPER*. Ces métriques ont été choisies pour le projet et comparées avec d'autres métriques d'évaluation automatique fondées sur les notions de score grammatical ou sémantique telles que *X-Score*, *D-Score* and *WNM*.

Les avantages et les implications des différentes métriques ont été longuement étudiées dans une première phase de méta-évaluation, avant l'évaluation humaine.

Pour la première campagne, cinq systèmes ont été évalués sur la tâche Anglais=>Français et deux systèmes ont participé à la tâche Arabe=>Français. Le corpus de test en anglais était composé de 15 documents issus du corpus *JOC* et le corpus en arabe de 16 documents provenant de 32ème Conférence Générale de l'Unesco.

Dans la deuxième campagne, 6 systèmes ont participé à la tâche Anglais=>Français et un seul système

à la tâche Arabe=>Français. Les corpus utilisés pour cette campagne étaient constitués d'une part d'un ensemble de textes médicaux provenant du projet CESART pour la première tâche, et d'un ensemble de textes extraits pour une part du site de l'Unesco et pour une autre part du journal *Al Hayat* pour la tâche Arabe=>Français. Dans les deux camapgnes, le protocole a consisté à noyer les corpus de test dans un plus grand corpus homogène sur le plan thématique.

3.2.4. EASy

L'objectif de la campagne **EASy** a été de concevoir et testé une méthodologie d'évaluation pour comparer des analyseurs syntaxiques du français et produire des ressources linguistiques volumineuses et validées. Celles-ci ont été produites en combinant automatiquement les corpus annotés produits par les systèmes.

Les corpus produits sont des textes relevant de différents genres (littérature, médecine, technique, langue générale, etc.) et provenant de différentes sources (journaux, questions, sites web, transcriptions de la parole, etc.). Les principaux résultats du projet sont un protocole complet d'évaluation comprenant la méthode de constitution des corpus, le manuel d'annotation du corpus, l'évaluation des analyseurs ayant participant à la campagne et un volume important de ressources annotées (*Treebank*).

Les mesures d'évaluation adoptées ont été les méthodes classiques de précision et de rappel qui ont été calculées sur les deux tâches : l'évaluation de constituants et l'évaluation des relations de dépendance. Néanmoins, dans la mesure où les corpus étaient

hétérogènes, les calculs ont été effectués par type de corpus. De plus, les résultats par type de corpus permettent de savoir quel parseur est le plus adapté à tel ou tel contexte.

Dans le même esprit, il a été considéré pertinent d'évaluer les différents sous-ensembles de relations de dépendance séparément. D'abord, les relations de base telles que sujet ou objet direct, ensuite les relations impliquant les modifieurs (de noms, d'adjectifs…) et enfin les relations les plus difficiles à calculer comme la coordination, l'apposition ou la juxtaposition.

Cinq fournisseurs de corpus ont participé à la campagne produisant ainsi un corpus de genres différents de 82 734 mots annotés. Cette partie a été « cachée » dans un corpus global de 769 154 mots. L'évaluation a été faite sur la partie annotée du corpus mais les 16 systèmes impliqués dans le processus ont dû annoter la totalité du corpus produisant ainsi un résultat indirect qui est donc l'annotaion d'un corpus plus volumineux.

3.2.5. EQueR

Le but du projet était de définir un cadre d'évaluation pour les systèmes de Questions-Réponses en français et d'organiser une campagne d'évaluation. Deux tâches de recherche automatique d'informations ont été proposées : la première était une tâche générale de recherche dans une collection de textes hétérogènes, principalement des articles de journaux, et la seconde était une tâche spécialisée dans le domaine de la médecine portant donc sur des textes médicaux. Deux corpus de questions ont été constitués, respectivement de 500 et 200, conçus pour correspondre à 4 types de

questions, « factuelles » (Quand Kurt Cobain est-il mort ?), de « définition » (Qu'est'ce qu'IBM ?), « affirmative/négative » (Est)ce que Jean-Paul II est allé en Chine ?) et des questions de « liste » (Quels sont les 7 pays qui forment le G7 ?).

La phase d'évaluation des systèmes s'est déroulée sur le site même des participants et a duré une semaine. A chaque question, le système devait répondre une courte réponse et/ou un paragraphe de 250 caractères. Les deux types de réponses étaient ensuite évaluées par deux juges humains.

Les principaux résultats du projet consistent en un corpus du français général (environ 1,5 GO), principalement des articles de journaux, un corpus du français médical (environ 40 G0) constitué de textes provenant de différents sites médicaux, un corpus de 500 questions générales et un corpus de 200 questions relevant de la médecine. Par ailleurs, un outil semi-automatique d'évaluation destiné à aider les juges humains à évaluer les réponses (l'outil *EvalQA* et sa documentation) est disponible ainsi qu'un outil d'évaluation automatique permettant d'évaluer les réponses et de calculer les mesures de performance sans intervention humaine.

3.2.6. ESTER

La campagne d'évaluation **ESTER** a visé l'évaluation des systèmes de transcriptions d'émissions radiophoniques. Les transcriptions ont été enrichies par un ensemble d'informations annexes, comme le découpage automatique en tours de paroles, le marquage des entités nommées, etc. L'évaluation de la qualité des

informations annexes en plus de l'évaluation de la transcription orthographique a permis d'établir une référence des niveaux de performances actuels de chacune des composantes d'un système d'indexation, tout en donnant une idée des performances du système complet.

L'évaluation a proposé trois classes de tâches : la transcription orthographique, la segmentation et l'extraction d'informations. Huit laboratoires ont participé à la tâche de transcription orthographique.

Les tâches de segmentation se sont décomposées en segmentation en événements sonores, suivi de locuteurs et segmentation en locuteurs. Pour la segmentation en événements sonores où la tâche a consisté à détecter les parties contenant de la musique (avec ou sans parole) d'une part et les parties comprenant de la parole (avec ou sans musique), sept laboratoires ont participé. Trois laboratoires ont participé à la tâche de suivi de locuteur qui a consisté à détecter les parties du document correspondant à un locuteur donné et quatre systèmes ont enfin participé à la tâche de segmentation en locuteurs qui consiste à segmenter le document en locuteurs et regrouper les parties parlées par le même locuteur.

Le dernier type de tâche, qui consistait à détecter des entités nommées, a été évalué de façon plus exploratoire que les tâches précédentes. Le but était de mettre en place et tester un protocole d'évaluation plutôt que de mesurer les performances. Les systèmes devaient détecter huit classes d'entités (personne, lieu, date, organisation, entité géo-politique, montant, bâtiment et inconnu) à partir de la transcription automatique ou de la

transcription manuelle. Trois systèmes ont participé à cette tâche exploratoire.

Les ressources produites dans le cadre du projet sont 100 heures d'émissions transcrites orthographiquement et annotées en entités nommées et 1 700 heures d'enregistrements d'émissions radiophoniques non transcrites. La phase d'évaluation a également permis de mettre à jour le corpus en analysant les sorties automatiques des systèmes et les transcriptions manuelles. Le dictionnaire d'équivalence a également été enrichi au terme de la phase d'adjudication. Le corpus de 100 heures de transcriptions inclut 1,2 millions de mots pour un vocabulaire de 37 000 mots. 74 082 occurrences d'entités nommées sont annotées pour un total de 15 152 entités différentes. Les ressources textuelles reposent essentiellement sur les archives du journal *Le Monde* et du corpus des débats du Conseil européen. Enfin, des guides et des manuels ont été produits et sont fournis dans le « package » distribué par ELDA.

3.2.7. EVASY

Le projet **EVASY** a été consacré à l'évaluation des systèmes de synthèse de la parole à partir du texte (TTS) en Français. Il a porté sur quatre types d'évaluation différente. La première s'est attachée à évaluer la qualité de conversion graphème-phonème (GP) sur la tâche de conversion des noms propres qui est considérée comme l'une des plus difficiles. La seconde évaluation a porté sur la prosodie. La troisième a porté sur un test d'intelligibilité, avec un nouvel ensemble de phrases sémantiquement imprédictibles. Enfin, la dernière

évaluation a concerné la mesure de la qualité globale de la parole synthétisée en utilisant un test d'opinion.

Quatre systèmes ont participé à la tâche de conversion GP qui a consisté à phonétiser une liste de noms propres en 3 heures. Pour réaliser cette tâche, une liste de 4 115 couples prénom-nom a été produite, extraite du journal *Le Monde* des années 1992–2000 (plus de 200 millions de mots) et transcrite manuellement dans l'alphabet phonétique SAMPA. Cette liste a ensuite été enrichie d'informations concernant l'origine des noms de famille.

La deuxième évaluation a concerné la prosodie selon une méthode permettant d'évaluer la prosodie indépendamment des autres modules de chaque système (comme les traitements textuels et la synthèse acoustique). Cinq systèmes ont participé à cette partie de la campagne dont 3 systèmes par diphones et deux systèmes par sélection/concaténation. Le corpus d'évaluation comprenait sept phrases phonétiquement équilibrées extraites du corpus BREF. Les phrases produites par les systèmes de synthèse ont ensuite été évaluées par des 19 assesseurs par rapport à une référence naturelle.

La troisième partie de l'évaluation a consisté à comparer la qualité des systèmes à base de diphones et celle des nouveaux systèmes par sélection d'unités.

Le test d'intelligibilité mis en œuvre a utilisé des phrases sémantiquement imprédictibles, *Semantically Unpredictable Sentences* (SUS), un paradigme expérimental permettant l'évaluation objective de l'intelligibilité au niveau des mots. Six systèmes de

synthèse à partir du texte en français, représentatifs du niveau actuel de qualité atteint, ont été mis à l'épreuve.

Trois systèmes utilisaient des diphones et trois étaient des systèmes par sélection/concaténation. Pour cette tâche, une liste de 288 SUS a été construite, contenant 4 types de structures syntaxiques. Les 288 phrases ont été prononcées par les 6 systèmes, ainsi que par un locuteur professionnel afin de donner une référence naturelle. Les équipes participantes devaient rendre les phrases de synthèse dans les heures suivant leur réception. Des assesseurs ont ensuite jugé l'intelligibilité des résultats.

Le dernier aspect de la campagne a consisté à confronter le test d'intelligibilité sur les SUS à un test plus global utilisant des échelles de catégorie absolue. Six catégories ont été retenues, en plus de l'opinion moyenne : compréhension, agrément, (non-) monotonie, naturel, fluidité et prononciation, qui sont issues d'une adaptation au français des critères d'évaluation proposés dans le projet *Vermobil*. Comme pour le test *SUS*, les participants devaient synthétiser des centaines de phrases dans de brefs délais. Le corpus de phrases était extrait d'EUROM 1, collecté dans le cadre des projets *Multext* et *Esprit 2589/SAM*.

3.2.8. MEDIA

Le but du projet **MEDIA** a été de définir et d'expérimenter un protocole d'évaluation pour les systèmes de dialogue oral homme-machine dans un contexte applicatif qui était celui de la demande de renseignements (plus particulièrement la réservation d'hôtels à partir de sites web). Dans ce cadre, chaque

système devait convertir vers une représentation sémantique commune sa propre représentation de la demande de renseignements.

Le protocole d'évaluation a donc comporté la définition et l'utilisation de batteries de tests issues de corpus réels, une représentation sémantique et des métriques communes. Ce paradigme a permis l'organisation de deux campagnes : une campagne d'évaluation de la compréhension hors contexte et une autre en contexte.

Cinq systèmes fondés sur des approches différentes ont participé à la première campagne qui s'est déroulée en deux phases. D'abord un test à blanc sur un corpus de 1 000 énoncés qui a permis la définition du protocole de test puis l'évaluation littérale qui a été menée sur un corpus de 3 000 énoncés. Deux systèmes ont participé à la seconde campagne.

Outre l'évaluation des systèmes, le projet a permis la production d'un corpus d'environ 70 heures de dialogues oraux intégralement transcrits et annotés sémantiquement. L'annotation du corpus permet de représenter le sens littéral des énoncés et s'appuie sur un dictionnaire sémantique correspondant au domaine de la tâche évaluée (la réservation d'hôtels).

3.3. Action de normalisation

Dans le cadre du « cluster » **NORMALANGUE**, deux projets ont visé à renforcer la participation française dans les instances de normalisation et de standardisation.

Le projet **RNIL** a permis à un consortium d'entreprises et de laboratoires de participer à l'élaboration des normes concernant la description des

ressources linguistiques pour l'écrit au sein du comité ISO TC37 SC4. Un comité national a été mis en place par l'AFNOR (*Association française de normalisation*) à la fois pour élaborer la position française et rendre compte de l'avancement des travaux. Un aspect important du projet concernait en effet la diffusion rapide des informations provenant du comité ISO vers les acteurs académiques et industriels. Les travaux ont principalement porté sur l'annotation morpho-syntaxique, le modèle de lexique et le registre des catégories de données. Compte tenu des délais propres au processus de normalisation, les travaux ne sont pas achevés mais des résultats significatifs ont déjà été obtenus.

Le second projet, **Technovox**, a concerné le domaine de la parole et s'est focalisé sur les instances de standardisation. L'évolution technologique très rapide du domaine confère en effet aux différents comités et forums industriels un rôle majeur dans l'élaboration de standards.

Les travaux ont concerné l'enrichissement des standards *VoiceXML* et *UNL*. Le projet a permis la participation à différentes instances telles que le Forum VoiceXML du W3C, la Fondation UNL, le comité AURORA de l'ETSI, le Forum SALT…

3.4. La mise en place d'un dispositf de veille

Enfin, le dernier objectif de *Technolangue* était de mettre en place un dispositif de veille dans le domaine des technologies de la langue. Celui-ci a pris la forme d'un portail dédié conçu en collaboration avec les deux sociétés savantes du domaine en France, l'ATALA (*Association pour le Traitement Automatique des Langues*) pour le domaine de l'écrit et l'AFCP

(*Association Française de la Communication Parlée*) pour la parole, ainsi que l'APIL (*Association des Professionnels des Industries de la Langue*) qui regroupe les industriels du secteur.

Le portail est accessible à l'adresse : www.technolangue.net et donne accès à des informations permettant de comprendre les enjeux stratégiques des technologies de la langue (brèves d'actualité, définition du secteur, annuaire des acteurs, informations économiques, présentation de projets…). Le portail est un élément du réseau d'information constitué par les sites web des partenaires, ATALA, AFCP et APIL.

5. Conclusion

En conclusion de cette rapide présentation des projets qui ont été menés dans le cadre de *Technolangue*, trois points saillants peuvent être rappelés. Tout d'abord, cette action interministérielle n'a pas été un programme classique de R&D mais a permis la création de différents types de ressources linguistiques pour la langue écrite et orale, visant ainsi à combler un manque identifié depuis longtemps par la communauté. A l'issue du programme, il serait prétentieux de considérer que l'objectif a été totalement atteint mais le retard a été en partie rattrapé. Les ressources produites sont en effet nombreuses et, dans l'ensemble, conformes à l'état de l'art en termes de balisage et d'annotation. Elles proviennent à la fois des projets qui visaient explicitement la production de ressources mais également, et pour une grande part, des campagnes d'évaluation. Les ressources provenant des huit campagnes d'évaluation sont multiples : corpus écrits et oraux annotés, corpus de requêtes et référentiels.

Associés aux ressources, les logiciels utilisés pour la mesure des performances sont également disponibles. Ces « boîtes » d'évaluation, qui sont diffusées par ELDA, permettent aux laboratoires et aux entreprises d'organiser leurs propres tests en interne.

Un deuxième aspect du programme a été de veiller aux conditions de diffusion des résultats, particulièrement pour les acteurs de la recherche académique. Ainsi, de nombreuses ressources sont diffusées gratuitement ou à un prix extrêmement bas (prix de reproduction du support) dans un cadre juridique autorisant leur libre utilisation à des fins de recherche. Un modèle unique de distribution n'a toutefois pas pu être imposé dans la mesure où différents projets ont été financés sur la base du coût partagé entre les ministères et les industriels.

Malgré quelques restrictions regrettables dans la diffusion des ressources, la plupart des acteurs ont joué le jeu contribuant ainsi à renforcer l'idée que la disponibilité de ressources de base (linguistiques et logicielles) est fondamentale pour le traitement automatique d'une langue. Autrement dit, pour qu'une langue soit une langue de communication à l'ère du numérique, pour qu'elle soit prise en compte par les applications de traitement automatique, il est indispensable que des ressources nombreuses et de qualité soient disponibles.

Un troisième aspect du programme est l'implication des laboratoires et des industriels dans les campagnes d'évaluation. Cet intérêt, qui se manifeste depuis longtemps pour les campagnes d'évaluation internationales comme CLEF (*Cross-Language Evaluation Forum*) et celles qui sont organisées aux

Etats-Unis comme TREC (*Text REtrieval Conference*), montre que les « offreurs de technologies » prennent conscience que l'évaluation est un aspect important de l'innovation technologique. Les discussions concernant les protocoles existants qui ont souvent mené à l'élaboration de nouvelles approches indiquent aussi que l'évaluation est un objet de recherche important, à la croisée de plusieurs disciplines.

Références

- CHAUDIRON S. (dir.), *L'Évaluation des systèmes de traitement de l'information*, Paris, Hermès, 2004.

- MERTENS P. *et al.* (dir.), *Verbum ex machina, Actes de la conférence TALN, Atelier Les Ressources dans le traitement de la langue écrite : L'apport de Technolangue et les enjeux industriels*, Presses universitaires de Louvain, Louvain-la-Neuve, 2006.

- Site Technolangue : www.technolangue.net qui permet d'accéder aux bibliographies des différents projets.