La lettre d'information

EUROPEAN LANGUAGE RESOURCES ASSOCIATION ELRA

Mai 1998

Vol.3 n.2

## Sommaire



*Les articles d'auteurs représentent le point de vue de leur signataire et ne reflètent pas nécessairement l'opinion des éditeurs ni la position d'ELRA/ELDA.*

***Directeur de la publication :***
Khalid Choukri

***Editeurs :***
Deborah Fry
Malin Nilsson

***Maquette :***
Rébecca Jaffrain

***Ont participé à ce numéro :***
*José Coch*
*Frances C. Johnson*
*Steven Krauwer*
*Anna O'Hora-Bimbot*
*Malin Nilsson*
*Vera Semenova*
*Richard H. Wojcik*

*ISSN: 1027-6564*

***ELRA/ELDA***
*Directeur Exécutif :*
*Khalid Choukri*
*Assistante : Rébecca Jaffrain*

*55-57, rue Brillat Savarin*
*75013 Paris - France*
*Tél. : (33) 1 43 13 33 33*
*Fax. : (33) 1 43 13 33 30*
*Courrier électronique :*
*elra@calvanet.calvacom.fr*
*WWW :*
*http://www.icp.grenet.fr/ELRA/home.html*

***Ce numéro a été réalisé avec le soutien de la Délégation Générale à la Langue Française***

# *Cher membres d'ELRA,*

Au fur et à mesure que nous avançons dans l'année 1998, ELRA peut à nouveau faire valoir les progrès substantiels qu'elle a accomplis dans de nombreux domaines. Les ventes de ressources du premier trimestre 1998 se sont avérées plus de 6 fois plus importantes que celles observées pour la même période de l'année 1997, avec toujours 90 % de ces chiffres correspondant aux organismes privés. La répartition géographique des acheteurs reste à peu près stable, avec plus de 65 % d'entre eux émanant d'Europe.

Nous avons fait l'acquisition de plusieurs ressources nouvelles, notamment un dictionnaire morphologique du bulgare fourni par l'Université de Plovdiv, un ensemble de dictionnaires bilingues de Translation Experts Ltd, une base de donnée de parole en russe (enregistrée par microphone) fournie par STC, et SIelex, un lexique phonétique de l'allemand de Siemens AG.

Nous sommes particulièrement fiers d'annoncer également la disponibilité de nouvelles ressources réalisées dans le projet SpeechDat-II. La première langue disponible est l'allemand : 1 000 locuteurs enregistrés sur réseau téléphonique fixe. Dans le but d'encourager l'usage de cette base, ELRA propose une offre spéciale d'un lot contenant la base SpeechDat(M), en complément de cette nouvelle base de données, le tout pour un prix spécial. Pour plus d'informations, veuillez visiter notre site Web ou contactez le secrétariat d'ELDA.

Nous avons également l'immense plaisir de souhaiter la bienvenue à quatre nouveaux membres qui ont rejoint ELRA depuis la publication de la dernière Lettre d'Information. Il s'agit de CLIPS-IMAG, France (Collège Parole), LOCUS Dialogue, Canada (Parole), Polderland Language & Technologies, Pays-Bas (Terminologie) et Voice Control Systems, USA (Parole).

Le nouveau membre du Conseil d'Administration d'ELRA est Helmi Sonneveld de Topterm, dont un portrait est présenté dans cette Lettre. Avec son expérience en terminologie, Helmi sera certainement une très bonne représentante pour ce collège au sein du Conseil.

Le rapport sur l'étude de marché d'ELRA, conduite durant l'année 1997, a été adressé aux participants et aux membres de l'Association et nous attendons avec impatience vos éventuelles réactions. Nous vous rappelons également que nous sommes prêts à distribuer, si vous le souhaitez, toutes les ressources dont vous disposez.

Maintenant que les travaux sur les procédures de validation sont terminés ou en passe de l'être pour les trois Collèges, nous allons nous consacrer à mettre sur pied des sites de validation pour tester les manuels et les appliquer à des situations réelles. Nous vous incitons à nous contacter si vous êtes susceptibles d'être intéressés pour participer à cette phase du travail, qui sera, comme la précédente, coordonnée par ELRA.

Par ailleurs, comme c'est le cas pour tous les projets du secteur de l'Ingénierie Linguistique, nous avons participé à la réunion de revue annuelle du travail réalisé dans les projets du secteur IL, laquelle a eu lieu à Mondorf, du 17 au 19 mars. Les résultats de la revue ont démontré l'efficacité du projet ELRA dans son travail d'intermédiaire entre les fournisseurs et les utilisateurs de Ressources Linguistiques, ainsi que le succès de ses activités de promotion et de dissémination. L'incertitude quant à la capacité d'ELRA d'être à la fois un forum d'intérêt public, et une société financièrement indépendante a conduit les experts à demander plus d'informations sur nos plans d'avenir.

Le programme détaillé de la conférence LREC, ainsi que d'autres informations utiles, vous sont proposés dans les pages centrales de ce numéro. Nous tenons à souligner tout particulièrement le grand nombre de workshops et de tables rondes organisés à cette occasion : ceux-ci tendent à promouvoir les débats et les échanges au sein de la communauté des sciences du langage.

Dans ce numéro de la Lettre d'Information d'ELRA, nous vous proposons des articles sur des sujets spécialisés, notamment un article sur le résumé automatique de textes, de Frances Johnson de la Metropolitan University à Manchester, et un autre sur les langages contrôlés, de Richard Wojcik de Boeing (un précurseur dans ce domaine). De plus, Steven Krauwer d'ELSNET, avec qui ELRA a signer un accord de coopération, plaide pour une plus grande disponibilité des ressources à travers l'Europe, notamment l'Europe de l'Est. Dans le prolongement de ce constat, Vera Semenova de SCIPER présente un panorama des ressources russes. Enfin, José Coch d'ERLI rend compte des travaux du projet MULTIMETEO.

Nous espérons que ces articles vous intéresseront, ainsi que les autres informations contenues dans cette Lettre. Nous souhaitons avoir l'occasion d'échanger avec vous nos points de vue de vive voix, lors de la conférence LREC à Grenade.

Antonio Zampolli, Président

Khalid Choukri, Directeur Exécutif

# Portrait

*Helmi Sonneveld*

Helmi Sonneveld (1959) est une terminologue confirmée, connue au niveau international dans le domaine de la terminologie, et auteur d'un grand nombre de publications sur le sujet. Elle est diplômée en "Langue anglaise et Linguistique" avec mention de l'Université Libre d'Amsterdam en 1985. Elle a alors travaillé à l'Université Libre pendant deux ans, où elle commença par donner des cours en terminologie et lexicologie avant de fonder le bureau de terminologie Topterm en 1987. Elle est actuellement à la direction de la Division Europe de Topterm, dont les bureaux sont établis à Amstelveen, Pays-Bas, et Columbus, Ohio, Etats-Unis. En 1995, Topterm a été un partenaire actif pour les Pays-Bas dans le projet POINTER, soutenu par l'Union Européenne.

Helmi Sonneveld a participé à un grand nombre de projets dans le domaine de la terminologie et des thesauri. Depuis 1990, elle a notamment travaillé dans le secteur médical en tant que conseillère au Conseil National Néerlandais de la Santé sur la traduction de l'ICD-10 (*International Classification of Diseases*, Classification internationale des maladies). D'autre part, elle a travaillé en tant que conseillère pour différents programmes soutenus pour l'Union Européenne dans le domaine de la médecine.

Helmi Sonneveld est également éditrice en chef d'un journal international sur la terminologie, intitulé *Terminology: International Journal of Theoretical and Applied Issues in Specialized Communication* (Terminologie : Journal international sur les problèmes théoriques et appliqués en communication spécialisée), publié par John Benjamins Publishing Company, Amsterdam/Philadelphie. En octobre 1997, elle fut élue Présidente de l'Association Européenne pour la Terminologie (EAFT) lors de la seconde Assemblée Générale, qui s'est tenue à Barcelone. En tant que Présidente de l'EAFT, Helmi Sonneveld est membre de plusieurs comités scientifiques travaillant sur la préparation de conférences et symposiums. Actuellement, elle organise un symposium scientifique sur les principes de la théorie terminologique.

Helmi Sonneveld est à l'origine de la création, en 1997, de l'Association terminologique néerlando-flamande (NL-TERM) dont elle est aujourd'hui vice-présidente.

# Rapport final sur l'étude de marché d'ELRA de 1997 – Extrait

*Malin Nilsson, ELRA*

L'année dernière, ELRA a réalisé deux études de marché dont les résultats ont été dépouillés et synthétisés dans un rapport qui est désormais accessible à nos membres et aux participants à l'étude. Plusieurs d'entre vous ont pris part, ou entendu parler d'au moins une de ces études. La première est connue sous le nom de l'Etude de Printemps et la seconde comme l'Etude des Besoins des Utilisateurs. Toutes les deux visaient à collecter autant d'informations intéressantes que possible en ce qui concerne les ressources linguistiques, qu'il s'agisse des utilisations, des besoins, du marché ou des développements présents et futurs de ces ressources, selon leur type. Cet article présente un bref compte rendu des résultats de cette enquête en se concentrant sur les chiffres relatifs aux besoins et aux acquisitions de ressources linguistiques.

Pour commencer, il faut constater que parmi les domaines d'activités et d'applications dont proviennent la majorité des acteurs, c'est celui de la reconnaissance de la parole qui apparaît comme le plus grand consommateur de ressources en parole. Pour les utilisateurs de ressources textuelles, les secteurs d'activités dominants sont la recherche documentaire, l'indexation de documents, l'extraction de terminologie et l'évaluation, les langages contrôlés et les systèmes de traduction automatique. En ce qui concerne les utilisateurs de ressources terminologiques, mentionnons les domaines suivants : recherche documentaire, indexation de documents, création de thesaurus et consolidation de bases de connaissances.

Au niveau des besoins spécifiques, les langues les plus fréquemment citées dans l'expression des besoins sont l'anglais et les autres langues principales de l'Europe. Pour les organisations travaillant sur la parole, cette catégorie représente 56 % des besoins totaux, alors que pour les organisations intéressées par des textes, elle constitue 63 %. Pour le secteur de la terminologie, ce chiffre atteint même 68 %. Cependant, on observe des demandes pour des langues telles que le japonais et l'hébreu. De nombreuses réponses expriment un besoin pour des ressources multilingues (ou du moins, polylingues).

Les domaines d'utilisation les plus cités lors des enquêtes sont pour la parole, soit la téléphonie, soit d'autres domaines spécifiques. Pour le texte, un grand nombre de domaines différents sont mentionnés, plus particulièrement les utilisations généralistes ou dans le domaine technique. Enfin, pour la terminologie, c'est aussi le domaine généraliste qui est le plus fréquemment cité, à coté du domaine économique.

Quand il s'agit d'acquérir des bases de données, la majorité des organisations interrogées répondent qu'elles se procurent les bases de données extérieures auprès d'organisations comme ELRA. L'option la plus fréquente est l'achat (37 %), puis la production interne (24 %). Parmi les autres options mentionnées, citons l'emprunt de bases de données, l'utilisation de bases de données déjà existantes en interne, l'acquisition par l'intermédiaire de projets européens, l'échange avec d'autres organisations, l'utilisation de données publiques (notamment via Internet) ou l'obtention auprès de partenaires ou de laboratoires universitaires. Les réponses ne diffèrent guère entre les 3 collèges, sur cette question.

Les objectifs fixés initialement à cette étude se sont avérés difficiles à atteindre. De nombreuses raisons en sont à l'origine, et notamment le faible taux de participation. Cependant, ces résultats ont permis à ELRA d'améliorer sa perception du marché et ils seront utilisés pour mettre au point les futures politiques de l'Association en ce qui concerne la commercialisation, la communication et l'amélioration des services offerts à nos membres. Parmi les débouchés de ce travail, notons aussi la comparaison des résultats et la coordination des études à venir, avec celles conduites par d'autres organisations, notamment dans le cadre du projet Euromap.

Veuillez envoyer vos commentaires à :
Malin Nilsson
ELRA/ELDA
Tél. : +33 1 43 13 33 33
Courriel : elra-elda@calva.net

# ELSNET et ELRA : un passé et un futur communs

*Steven Krauwer (UiL OTS, Université d'Utrecht, Coordinateur d'ELSNET)*

*Dans cet article nous proposons une brève présentation d'ELSNET et de ses activités, nous mettons en évidence les intérêts communs d'ELSNET et d'ELRA et nous traçons les contours d'une coopération possible entre les deux organisations dans le domaine des ressources linguistiques. Cette action conjointe, dénommée BLARK, doit être lancée à l'occasion du Cinquième Programme Cadre de la Commission Européenne.*

## Objectifs d'ELSNET

ELSNET, Réseau d'Excellence Européen sur la Langue et la Parole (European Network of Excellence in Language and Speech) a été créé en 1991, en même temps que deux autres réseaux pilotes soutenus par le Programme de Recherche à Long Terme ESPRIT (LTR) de la Commission Européenne. Le réseau est hébergé par l'Institut de Linguistique (OTS) à l'Université d'Utrecht. ELSNET s'intéresse plus particulièrement aux problèmes liés au développement de systèmes multilingues intégrés pour le traitement du langage parlé et écrit à couverture illimitée. ELSNET regroupe les principales équipes de recherche européennes dans le domaine du traitement du langage écrit et parlé. Il recense actuellement 50 membres du milieu industriel et 80 membres du monde académique. Une liste complète des membres d'ELSNET peut être consultée sur : http://www.elsnet.org/about-elsnet/organised/elsnetNodes.html

Les activités d'ELSNET visent à favoriser, soutenir et coordonner les efforts de ses membres dans la mise en œuvre de systèmes de traitement du langage écrit et parlé. L'objectif d'ELRA est de devenir le centre de collecte et de distribution de ressources linguistiques tant pour l'oral, que pour l'écrit ou la terminologie.

Une attention particulière est apportée à l'intégration de l'oral et de l'écrit car il apparaît que les deux communautés tendent à évoluer de part et d'autre d'un fossé culturel et méthodologique qui tend à les séparer.

Depuis 1991, ELSNET a déployé ses activités dans 4 grands domaines : la formation, la coordination de la recherche, la dissémination de l'information et les ressources linguistiques.

## Formation

L'Ecole d'Eté Européenne d'ELSNET est devenue, au fil des ans, un événement apprécié qui attire les étudiants intéressés par les sujets situés au confluent du langage écrit et du langage parlé, tels que la prosodie, les méthodes à base de corpus, la multilingualité, les systèmes de dialogue et le développement de lexiques. En 1998, l'Ecole d'Eté aura lieu à Barcelone et sera consacrée à la robustesse. (http://gps-tsc.upc.es/veu/ess98).

Des cours intensifs, de durée limitée, traitant de sujets récents et pointus, ont pour vocation de maintenir les acteurs du milieu industriel informés sur les derniers développements dans des domaines spécialisés (systèmes de dialogue oral en 1997, terminologie en 1998-1999).

Par ailleurs, une action a récemment été mise en place pour développer un programme d'Etudes Supérieures Européennes en Traitement du Langage Ecrit et Parlé (European Masters Programme in Language and Speech), lequel a été initié par le Réseau Thématique Socrates intitulé « Sciences de la Communication Parlée » (http://tn-speech.essex.ac.uk/tn-speech).

## Coordination de la recherche

La coordination de la recherche est une question délicate, d'autant plus qu'ELSNET n'est pas un organisme chargé de financer les projets de recherche. Pour cette raison, les actions de recherche d'ELSNET sont indirectes et elles se concentrent sur l'amélioration des conditions dans lesquelles les membres de la communauté scientifique peuvent rapprocher, comparer et recouper leurs résultats. L'évaluation et la normalisation occupent une place primordiale dans nos priorités : plusieurs projets dans ces domaines ont été retenus, par exemple le projet DISC (ESPRIT-LTR / http://www.elsnet.org/disc), qui vise à mettre en place des recommandations concernant les systèmes de dialogue oral. Citons aussi le projet ELSE (LE-4) qui a pour objectif la création d'une infrastructure européenne pour l'évaluation dans le domaine de la parole et du langage écrit.

## Dissémination de l'information

ELSNET s'attache à tenir ses membres informés de toutes les activités et les actions liées au domaine par l'intermédiaire d'une lettre d'information bimensuelle, ELSNews (abonnement gratuit), d'un site Web (http://www.elsnet.org) et d'une liste de diffusion électronique (elsnet-list@let.ruu.nl). De plus, des services spéciaux pour la dissémination sont offerts aux projets rattachés à ELSNET.

## Notre passé commun : les ressources

Les ressources linguistiques ont toujours été, et continuent d'être, un aspect essentiel des activités d'ELSNET. Un groupe de travail spécial au sein d'ELSNET est chargé de mettre au point des initiatives sur ce sujet. Pour ceux qui n'ont pas été impliqués dans ELRA depuis ses débuts, il faut signaler qu'ELRA et ELSNET ont un passé commun : le projet RELATOR qui a été le point de départ de la création d'ELRA et qui a été lancé à l'initiative d'ELSNET. Ceci souligne clairement l'intérêt d'ELSNET pour les ressources linguistiques lequel n'a pas diminué depuis la création. Bien que la constitution de ressources nouvelles soit bien au-delà des possibilités financières d'ELSNET (et ne rentre pas dans le cadre de notre contrat avec la Commission Européenne), nous explorons en permanence de nouveaux types de données et de nouvelles manières de les annoter. Des études pilotes, conduites sous l'impulsion de notre groupe de travail sur les ressources ont notamment consisté à annoter une partie des données du CDROM ECI en allemand et en italien, en utilisant les recommandations d'EAGLES sur l'annotation. Actuellement, nous effectuons aussi des expériences en annotation sémantique. Par ailleurs, le projet MATE (LE-4) qui vient de débuter, a pour but de développer des méthodes et des outils pour l'annotation de dialogues. ELSNET et ELRA ont en outre mis au point un accord formel de collaboration qui a débouché sur la distribution par ELRA de données produites par ELSNET, ainsi que sur des actions communes à l'occasion d'Eurospeech 1997 et une collaboration étroite à l'occasion de nouveaux projets.

## Notre futur commun

On peut déduire de ce qui précède qu'ELSNET et ELRA ont devant eux un futur commun très intéressant. Mais il y a un domaine, extrêmement exaltant d'après nous, où aucune collaboration n'a été établie pour l'instant, mais où ELSNET et ELRA peuvent prendre plusieurs initiatives essentielles dans le futur proche : la coopération avec les pays de l'Europe Centrale et de l'Europe de l'Est.

L'horizon géographique d'ELSNET s'étend au-delà des frontières de l'Union Européenne et ELSNET s'est intéressé à l'Europe Centrale et à l'Europe de l'Est dès 1994. Grâce à un financement supplémentaire de la Commission Européenne, une première étude a été effectuée pour identifier les principaux acteurs dans le domaine du traitement de la parole et du langage naturel en Europe Centrale et en Europe de l'Est. Bien que cette étude n'avait pas la prétention d'être exhaustive, elle a permis de donner des informations plus précises à la communauté scientifique occidentale sur cette vaste région géographique.

Dans le même temps, ELSNET a obtenu l'autorisation nécessaire pour permettre à quatre laboratoires de premier plan, issus de Hongrie, de la République Tchèque, de Roumanie et de Bulgarie, d'adhérer à son réseau de membres.

Le projet « ELSNET va vers l'Est » (ELSNET goes East) constitue une troisième initiative à l'intention des ex-Pays de l'Est. Ce projet, qui s'est déroulé de 1995 à 1997, résultait d'une action concertée dans le cadre du programme Copernicus et visait à mettre en place les bases d'une extension d'ELSNET aux pays d'Europe Centrale et d'Europe de l'Est (http://www.wins.uva.nl/research/illc/ege/elsea

st.html). Il a été largement couronné de succès et a abouti à une situation où pas moins de 12 membres d'ELSNET proviennent de l'Europe Centrale ou Orientale (dont 3 PME). De plus, une nouvelle étude sur les principaux acteurs de ces pays, toujours dans le domaine du traitement de la parole et du langage naturel, a été réalisée et est sur le point d'être publiée (vous trouverez l'URL correspondante sur le site Web d'ELSNET).

C'est un des buts d'ELSNET de mettre sur pied un réseau pan-européen au sein duquel tous les acteurs du domaine pourront participer aux futurs développements des systèmes de traitement de l'Ecrit et de l'Oral, sur un pied d'égalité.

A l'heure actuelle, très peu de nos collègues d'Europe Centrale ou d'Europe de l'Est participent à des projets européens ou sont à même d'entrer en compétition scientifique ou commerciale avec des organisations ou des institutions de l'Europe de l'Ouest. Un des facteurs limitants est indiscutablement le manque de ressources financières. Même dans les pays sur le point de rejoindre l'Union Européenne, les conditions demeurent très difficiles et n'évolueront que très progressivement. Ni ELSNET ni ELRA ne sont en mesure de remédier à cette situation, mais il est essentiel de garder cette réalité présente à l'esprit lors de nos contacts et de nos échanges avec nos collègues de ces pays.

Ceci étant dit, la situation financière n'est pas la seule difficulté : en général, les chercheurs de l'Est sont désavantagés en termes d'accès aux connaissances et à l'expertise.

A ce niveau, il me semble que des organisations comme ELRA et ELSNET peuvent jouer un rôle significatif, au même titre que les sociétés savantes actives dans le domaine de la Communication Parlée ou du Langage Naturel, comme l'ESCA ou l'EACL, par exemple.

Dans le domaine des ressources linguistiques, sujet sur lequel les intérêts d'ELSNET et d'ELRA se recoupent, je pense qu'il serait souhaitable que le Cinquième Programme Cadre soit l'occasion d'au moins une action d'envergure qui bénéficierait non seulement aux pays de l'Europe Centrale et de l'Europe de l'Est, mais également aux pays dont les langues sont peu répandues. Voici les contours possibles d'une telle action :

Etape n° 1 : nous définissons de façon tout à fait générique et applicable à toutes les langues, un BLARK (Kit de ressources linguistiques de Base) contenant :

(a) Un corpus de texte généraliste minimal nécessaire pour tous travaux de recherche précompétitive dans cette langue, par exemple 10 millions de mots issus de journaux récents, annotés selon des normes communément admises.

(b) Un corpus similaire pour le langage parlé.

(c) Un ensemble d'outils de base permettant de manipuler et d'analyser ces corpus.

(d) Un ensemble de techniques et de compétences qui constituent le point de départ minimal pour le développement d'un secteur industriel compétitif sur les technologies du traitement du langage écrit et parlé.

Etape n° 2 : Nous évaluons pour chaque langue dans quelle mesure le BLARK correspondant existe déjà et quelles sont les pièces manquantes. Il peut s'agir de données, d'outils, ou de techniques et de compétences spécialement adaptés à une langue précise ou à une communauté linguistique particulière. Ce système s'applique également aux aspects multilingues.

Etape n° 3 : Nous mettons en place un ensemble d'actions coordonnées permettant de combler les lacunes observées, en préparant des propositions de projets visant à fournir la couverture minimale pour les points (a) à (d) listés ci-dessus, pour toutes les langues européennes.

**Trois questions viennent immédiatement à l'esprit :**

**1) Cette action présente-t-elle une utilité ?**

Notre réponse est oui. Nous ne devons certes pas nous bercer d'illusions en feignant de croire que le BLARK en tant que tel constituerait un point de départ suffisant pour mettre au point de vraies applications commerciales. Par contre, ce que le BLARK devrait permettre, c'est une base suffisante pour des travaux de recherche exploratoire, pour le développement de démonstrateurs pilotes et, ce qui n'est pas l'objectif le moins important, pour la formation d'une nouvelle génération de jeunes chercheurs et ingénieurs de développement.

**2) Qui devrait financer un telle action ?**

Je pense qu'il est clair que la Commission Européenne peut jouer un rôle essentiel à ce niveau, en étroite collaboration avec les instances nationales. On peut objecter que le fameux « principe de subsidiarité » inciterait à se tourner vers les gouvernements nationaux plutôt que vers la Commission, mais je vois au moins 3 raisons pour lesquelles cette action devrait être engagée au niveau européen :

(i) Tout d'abord, le coût d'une telle action est élevé,

(ii) Ensuite, il faut garder présent à l'esprit que le mode de financement des projets par la Commission, qui met en général l'accent sur l'intérêt des industriels et des utilisateurs, a tendance (sciemment ou non) à favoriser les trois ou quatre langues européennes les plus intéressantes commercialement, et il serait donc tout à fait légitime que la Commission accorde un soutien spécial aux autres langues, qui ne bénéficient pas d'un important lobbying industriel,

(iii) Enfin, une organisation de cette action à l'échelle européenne permettrait de réduire les coûts, à la fois au niveau financier et au niveau intellectuel, en offrant la possibilité de développer des synergies entre équipes de pays différents confrontés aux mêmes problèmes. L'aspect multilingue de ce projet implique un soutien crucial de la part de la CCE.

**3) Quel pourrait être le rôle d'ELSNET et celui d'ELRA ?**

Je crois que la réponse est simple : ELSNET et ELRA bénéficient d'une expérience et d'une expertise considérables, et disposent de précieuses ressources linguistiques et intellectuelles. Celles-ci ne sont peut-être pas immédiatement généralisables à toutes les langues, mais, si elles sont correctement partagées et diffusées, elles peuvent fournir un excellent point de départ pour l'action envisagée. Ceci éviterait d'aboutir à une situation où l'on « réinvente la roue » sans cesse, pour chaque langue nouvelle. De plus, la présence d'ELSNET et d'ELRA dans cette action garantirait une certaine cohérence et une connectivité entre les résultats obtenus, ce qui permettrait de s'appuyer par la suite sur des synergies plus efficaces.

## Conclusions

Un BLARK pour chaque langue d'Europe, comme je le propose dans cet article, n'est pas une idée radicalement nouvelle, au sens où d'autres y ont déjà pensé. Pour certaines langues, il existe sûrement déjà, pour d'autres, l'essentiel des éléments est peut-être déjà en place. Il y a même eu un certain nombre de projets sur fonds publics (notamment dans le cadre des programmes Copernicus et INTAS) qui visaient à constituer des fragments de BLARK pour certaines des langues de l'Europe Centrale et de l'Europe de l'Est. Cependant, je pense qu'aucun appel n'a encore été lancé à la Commission Européenne et aux instances nationales pour qu'elles initient une action concertée, à grande échelle, afin de faire prendre corps à cette idée, et j'espère que cet article incitera au moins une partie des membres d'ELSNET et d'ELRA à nous aider à faire avancer les choses dans cette direction. Espérons qu'avant la fin du Cinquième Programme Cadre, chaque langue en Europe, à l'intérieur ou non de l'Union Européenne, disposera de son propre BLARK.

Steven Krauwer
UiL OTS, Utrecht University
Tél. : +31 30 253 6050
Fax : +31 30 253 6000
Courriel : s.krauwer@let.ruu.nl
http://www-sk.let.ruu.nl

**Il est important que les membres et les fournisseurs d'ELRA puissent aider ELRA et ELSNET à la collecte de Ressources Linguistiques qui constituent le kit de base pour le traitement automatique. ELRA et ELSNET lanceront des initiatives afin d'assurer qu'un tel kit puisse être finalisé aussi bien pour les Ressources Linguistiques monolingues que multilingues.**

# Sur les ressources russes de l'Ingénierie de la Langue

*Vera Semenova, ANALIT*

Le présent papier se réfère au *Répertoire des acteurs et des produits des industries de la langue dans les pays de l'ex-URSS*[1] qui a été réalisé à la demande du Ministère de la Recherche français, par la société russe ANALIT et la société française SCIPER de 1994 à 1997. L'objectif de ce travail était de permettre une meilleure connaissance des ressources et des acteurs russes par les acteurs français et de contribuer ainsi à établir des coopérations dans le secteur de l'ingénierie linguistique. On espère que l'information accumulée dans le Répertoire est aussi d'un grand intérêt pour les acteurs des autres pays.

Le premier tome du Répertoire présente les descriptions de 99 équipes. Le deuxième tome est constitué de chapitres regroupant les produits par catégories : systèmes documentaires, gestionnaires de terminologies et de dictionnaires, correcteurs orthographiques, TAO, analyseurs linguistiques, OCR, systèmes de synthèse de parole, systèmes de reconnaissance et de traitement de parole, systèmes de dialogue oral, etc.

Dans le chapitre « Ressources linguistiques » sont réunies les descriptions de ressources de la langue russe (et d'autres langues) disponibles sous forme électronique : des dictionnaires, des bases de données lexicales et terminologiques, des corpus de textes, des corpus de parole, des données phonétiques, etc.

Faute d'informations validées, le Répertoire n'est pas exhaustif en ce qui concerne les ressources russes. Une partie essentielle de celles-ci a été créée à l'époque soviétique, au sein des instituts académiques, alors que les questions de rentabilité n'étaient pas d'actualité. D'un côté, cela a permis d'accumuler des collections énormes de textes, de terminologies, de données phonétiques, etc., mais d'un autre côté, leurs formats particuliers n'étaient souvent destinés qu'à un usage interne. De plus, pour ce type de ressources les droits d'auteur restent le plus souvent un sujet de polémique.

Mais pendant la dernière décennie, le marché du logiciel est né et s'est développé, et actuellement on trouve aussi des RLs commercialisés par des sociétés privées.

En Russie, le passage de l'informatique « lourde » à la micro-informatique prit un retard important par rapport aux pays occidentaux. Chronologiquement il a coïncidé avec d'autres grands bouleversements dont la Perestroïka fut l'origine. L'un d'eux a été l'abaissement brusque et considérable des crédits pour la recherche. De ce fait, un grand nombre de projets académiques ont été arrêtés, les résultats ont été soit détruits, soit stockés sur bandes magnétiques.

On peut dire que parmi les systèmes de TAO, seuls les couples de langues anglais-russe et russe-anglais (sauf exception) ont survécu, les dictionnaires des autres couples de langues étant restés sur des unités périphériques des anciens ordinateurs.

A cette époque même les meilleurs systèmes de TALN avaient des dictionnaires relativement limités bien que les algorithmes étaient suffisamment évolués. Les dictionnaires créés à l'aide de traitement automatique des corpus de textes sont peu nombreux en Russie, la plupart de RLs pour les systèmes de TALN ont été établis manuellement, par les mêmes chercheurs que ceux qui développaient des logiciels.

En même temps (et jusqu'à aujourd'hui) il y avait des constructeurs de dictionnaires qui faisaient leur travail par des méthodes lexicographiques manuelles, sans l'assistance d'un ordinateur, en disposant leurs données sur des fiches papiers stockées dans des tiroirs. C'est de cette manière qu'ont été créés (et puis édités sous forme de livre ) les dictionnaires principaux du russe, et aussi plusieurs dictionnaires bilingues, comme le « Nouveau grand dictionnaire anglo-russe » en 3 volumes, sous la direction de l'académicien Yu. Apresyan et E.Mednikova (1993) qui couvre plus de 250 000 mots et qui est le plus complet parmi les dictionnaires anglais-russes.

Le Département du Fonds Computationnel de la Langue Russe (à l'Institut de la Langue Russe à Moscou) a joué un rôle très important dans le développement de l'ingénierie de la langue en Russie ; avant l'avènement de la micro-informatique, il a mis sous forme électronique les principaux dictionnaires de la langue russe et en premier « Le dictionnaire grammatical de la langue russe » de A. A. Zaliznyak contenant plus de 80 000 entrées. L'auteur de ce fameux dictionnaire a inventé des formalismes efficaces pour décrire la flexion du russe. Cette ressource est maintenant utilisée par la plupart des analyseurs linguistiques.

La mise sous forme électronique du dictionnaire de A. A. Zaliznyak s'est passée à l'époque où il n'existait pas encore de scanners. Elle a été réalisée par saisie manuelle, à partir du clavier. De la même façon ont été mis sous forme électronique d'autres dictionnaires classiques du russe n'existant d'abord que sous forme de livre : dictionnaire orthographique (édité par S. I. Barkhudarov, 90 000 entrées), dictionnaire de dérivations de A. N. Tikhonov (145 000 entrées), dictionnaire syntaxique de G.A. Zolotova et autres, et aussi un grand nombre de textes sources d'éminents écrivains et poètes russes des XIX-XXèmes siècles, un corpus de textes de média, un corpus de textes de parole retranscrite, etc.

Les temps ont changé, et quand en 1995 la société Medialingua (Moscou) s'est chargée de mettre sous forme électronique les plus importants dictionnaires bilingues réunissant le russe avec les langues de l'Europe de l'Ouest, elle s'est servi d'un scanner, d'un outil OCR et d'un logiciel évolué, développé spécialement pour la correction, le balisage et la gestion de dictionnaires électroniques bilingues.

En 1996, cette société à produit un CD-ROM contenant le « Nouveau grand dictionnaire anglo-russe » mentionné ci-dessus. Puis elle a mis sous forme électronique plusieurs autres dictionnaires anglais-russes et les dictionnaires allemand-russe et russe allemand de E. M. Rymashevskaya qui contiennent 44 000 entrées ce qui correspond à 120 000 mots allemands et à 130 000 mots russes.

Actuellement pour la scannérisation de dictionnaires bilingues, on utilise l'outil OCR nommé FineReader, produit et commercialisé par la société BIT (Moscou). Ce logiciel est bien adapté à la reconnaissance de textes bilingues, il est monofonte avec les moyens d'apprentissage qui permettent d'apprendre des fontes inconnues de tous types. De plus, FineReader possède un correcteur d'orthographe intégré qui sait reconnaître les langues et corriger des textes bilingues.

Les dictionnaires mentionnés ci-dessus portent sur le vocabulaire de base, tandis qu'il existe plusieurs ressources terminologiques. Une équipe à l'institut VNIIKI (Moscou) gère RosTerm - banque nationale de données terminologiques. Cette banque est multilingue (les équivalents anglais sont tous indiqués, les équivalents français et allemands parfois), elle

contient 180 000 termes provenant de plusieurs domaines.

En même temps, de grandes ressources terminologiques ont été accumulées par des instituts informatiques (par exemple, ATOMINFORM gère la terminologie anglais-russe du domaine nucléaire) et aussi par des bibliothèques, sous forme de classifications et de thésaurus utilisés pour la recherche d'information. Ainsi, la bibliothèque médicale centrale (Moscou) gère un thesaurus médical bilingue (17 000 articles) - version russe du thesaurus MeSH (Medical Subject Headings) ; la bibliothèque centrale d'agriculture gère la terminologie du domaine de l'agriculture et aussi de l'industrie de l'alimentation, pêcherie, gestion de forêts, protection de l'environnement etc. Leur thesaurus contient les termes russes et leurs équivalents en anglais et en latin.

Pour la gestion de ressources terminologiques, les organismes publics utilisent, dans la plupart des cas, le logiciel micro ISIS. En revanche, les sociétés privées qui produisent des dictionnaires électroniques, préfèrent utiliser leurs propres logiciels qui sont habituellement constitués d'une partie producteur - permettant de développer des dictionnaires - et d'une partie utilisateur qui est fournie avec le dictionnaire et qui permet de le consulter (par exemple, sur un CD-ROM). La société BIT produit les dictionnaires sous le format du gestionnaire Lingvo, qui est aussi utilisé par d'autres équipes comme un outil pour produire et distribuer des dictionnaires. La société ETS, producteur des dictionnaires bilingues non seulement pour les couples anglais-russe et allemand-russe, mais aussi pour les couples composées du russe et d'une langue scandinave, utilise Polyglossum.

La société Russicon (St. Petersbourg) a développé, par ses propres techniques, plusieurs dictionnaires de la langue russe, des autres langues de l'Europe de l'Est, et aussi des dictionnaires bilingues pour les paires composées d'une de ces langues et du russe. Les dictionnaires de cette société servent de base pour l'analyseur linguistique Russicon et du système d'EAO de langues.

Comme on l'a déjà dit, les dictionnaires construits par traitement automatique de corpus de textes, sont peu nombreux en Russie. Pour en donner des exemples, notons les dictionnaires monolingues crées par A. V. Trusov (120 000 racines pour le russe, 100 unités lexicales pour l'anglais, 96 000 formes de mots pour l'ukrainien) à l'aide du traitement de grands volumes de textes, selon une méthodologie originale. Ces dictionnaires sont utilisés par la famille des correcteurs orthographiques GLAGOL (du même auteur) qui sont intégrés dans plusieurs OCRs, traitements de textes, etc.

Un autre exemple est fourni par le département des études linguistiques à l'institut VINITI (Moscou) qui a développé les dictionnaires russe-anglais et anglais-russe pour les systèmes de TAO, RETRANS/ERTRANS, par le traitement interactif de grand volume des textes disponibles en versions parallèles russe et anglaise (corpus de textes de plus de 80 millions d'occurrences). Chacun des dictionnaires a plus de 600 000 entrées, dont plus de 75% sont des expressions composées de 2 à 17 mots.

Il ne faut pas oublier que la Russie n'est pas uniquement habitée par des russes, il y existe plusieurs autres populations, avec leurs propres langues maternelles. On peut citer les ressources électroniques de la langue tatare, y compris un dictionnaire orthographique du tatare utilisé par TATCOR - correcteur d'orthographe de textes tatares, et aussi des ressources phonétiques pour le système de synthèse de parole tatare, qui sont développés par le Laboratoire des problèmes d'intelligence artificielle (ville de Kazan).

Le Fonds Phonétique de la langue russe est développé par le département de phonétique à l'Université d'Etat de Saint-Petersbourg, en coopération avec l'Université de Ruhr (Allemagne). Le Fonds est conçu comme la réunion du matériel acoustique (toutes les formes et unités significatives pour le russe), des logiciels pour son traitement et analyse et des résultats de cette analyse. En particulier, la « Phonothèque des unités sonores » fait partie du Fonds Phonétique de la langue russe. Elle est constituée de syllabes, de mots et de textes. Les syllabes sont toutes les combinaisons possibles des consonnes avec les voyelles. La partie textuelle est constituée du texte contenant toutes les chaînes sonores les plus fréquentes pour le russe.

Le même département a produit un dictionnaire de diphones pour le système RUSVOX de synthèse de parole (développé en coopération avec le CNET (France)). Il contient 2537 unités - tous les diphones possibles en russe.

Une autre ressource phonétique importante est le dictionnaire des unités phonétiques du russe et, partiellement, d'autres langues, développé par le laboratoire de phonétique et de communication orale à la faculté philologique de l'Université d'Etat de Moscou. Cette ressource réunit plus que 100 000 unités de dimensions diverses - allophones, syllabes, mots, structures rythmiques, syntagmes, phrases, textes. Il est utilisé dans les applications portants sur l'évaluation des canaux radio-téléphoniques et pour le diagnostic des pathologies.

A la même faculté, le groupe phonétique auprès de la chaire de linguistique théorique et appliquée développe le Dictionnaire Prononçable du russe. A partir du dictionnaire de Zaliznyak et d'autres sources, on a construit un dictionnaire de toutes les formes de mots contenant, pour chaque forme de mot, sa transcription phonétique, y compris l'information prosodique et segmentale, et aussi les indications sur les variantes possibles de la prononciation des fragments de mot.

Pour la construction du dictionnaire, on a utilisé le transcripteur automatique faisant partie du système AGAFON de synthèse de parole. Pour AGAFON, le groupe a aussi développé une base de 667 allophones constituant l'ensemble d'éléments suffisant pour la synthèse de parole russe.

Il reste encore beaucoup de ressources décrites dans le Répertoire, qui non pas pu être présentées ici à cause de la longueur limitée de l'article. Ceux qui sont intéressés pour en savoir plus ou pour obtenir le Répertoire peuvent contacter la société SCIPER :

SCIPER
42 rue Paul Claudel
91000 Evry
France
Tél. et fax. : (+33 1) 69 78 32 61
Courriel : 101376.156@compuserve.com

Note :

1. La description des ressources linguistiques russes ne saurait être complète si l'on ne mentionnaient les équipes des pays de l'ex-URSS, qui ont travaillé pendant des années sur le traitement du russe et ont établi des coopérations étroites avec les équipes russes. Des équipes ukrainiennes et biélorusses et une équipe ouzbek ont donc été intégrées dans le Répertoire ; d'autres équipes mériteront d'y figurer dans une prochaine version.

# Première conférence internationale sur les ressources linguistiques et l'évaluation

*Grenade, Espagne, 28-30 mai 1998*

La Première Conférence Internationale sur les Ressources Linguistiques et l'Evaluation est organisée à l'initiative d'ELRA en collaboration avec d'autres associations et consortiums.

La conférence abordera les thèmes suivants : la disponibilité et les méthodes d'évaluation des ressources linguistiques, les technologies et produits pour la langue écrite et orale. Discuter de telles questions sur une base de coopération internationale peut apporter des avantages mutuels significatifs. Le but de cette conférence est de proposer un panorama de l'état de l'art, de discuter des problèmes rencontrés et des solutions possibles, d'échanger des informations sur les activités en cours ou en prévision, de présenter les ressources linguistiques et leurs applications, de débattre des méthodologies d'évaluation, de faire connaître les outils disponibles et d'explorer les possibilités de promouvoir des coopérations internationales.

## WORKSHOPS & SALON

Comme vous pouvez le lire dans le programme qui suit, huit workshops d'une-demie journée chacun précèdent la conférence les 26 et 27 mai. Deux workshops sont présentés en parallèle lors des séances du matin et deux sont également présentés en parallèle lors des séances de l'après-midi. La liste des workshops est la suivante :

- Coréférence linguistique - 26 mai, séance du matin
- Adaptation des ressources lexicales et des corpus aux sous-langages et aux applications - 26 mai, séance du matin
- Réduire l'effort pour l'acquisition de ressources linguistiques - 26 mai, séance de l'après-midi
- Evaluation des systèmes d'analyse syntaxique - 26 mai, séance de l'après-midi
- Vers une infrastructure d'évaluation européenne pour le langage naturel et la parole - 27 mai, séance du matin
- Les ressources linguistiques pour les langues minoritaires européennes - 27 mai, séance du matin
- Développement de bases de données orales pour les langues du centre et de l'est de l'Europe - 27 mai, séance de l'après-midi
- Distribution et accès aux ressources linguistiques - 27 mai, séance de l'après-midi

De même, un workshop de 2 jours se tiendra après la conférence, les 31 mai et 1er juin, et est intitulé "La gestion d'information multilingue : niveaux actuels et capacités futures".

Un salon industriel est organisé par ELRA. Ce salon se déroulera conjointement à la conférence et permettra à des sociétés et autres projets de présenter et promouvoir leurs produits et prototypes en ingénierie linguistique.

## TABLES RONDES

Table ronde des agences de financement : Les membres des principales agences finançant la recherche et le développement en ingénierie linguistique (NSF, ARPA, CE, etc.) discuteront des priorités et perspectives pour la coopération internationale.

• *Présidence :* Antonio Zampolli (ILC). • *Participants :* Roberto Cencioni (EC), Ron Larsen (ARPA), Gary Strong (NSF). • *Intervenants :* Nuria Bel (FBG), Ralph Grishman (NYU), Nancy Ide (Vassar College), Joseph Mariani (LIMSI), Nick Ostler (Linguacubun).

Table ronde sur la coopération entre l'Union Européenne et les autres pays dans le domaine des ressources linguistiques et l'évaluation :

• *Présidence :* Alain Servantie (DG XIII-INCO). • *Participants :* Eva Hajicova (Univ. Charles), Dan Tufis (Academie roumaine), Klara Vicsi (Univ. technique de Budapest), Mohamed Chad (Université de Fez), Salem Ghazali (IRSIT, Tunis), Daniel Martin Mayorga (Telefónica Argentine)

Table ronde d'Eagles sur les standards lexico-sémantiques pour les systèmes d'information : Ce groupe discutera des directives pour la standardisation du codage lexical avec une attention particulière sur les besoins en traduction automatique et systèmes d'information.

• *Présidence :* Antonio Sanfilippo (Sharp). • *Participants* : Nicoletta Calzolari (ILC), Patrick Saint-Dizier (IRIT), Piek Vossen (Univ. Amsterdam), Robert Gauzauskas (Univ. Sheffield), Sophia Anianadou (Univ. Manchester Metropolitan).
• *Intervenants :* Eduard Hovy (USC), Ralph Grishman (NYU), Sergei Nirenburg (EMU).

Table ronde sur l'usage industriel des ressources linguistiques : Les utilisateurs et fournisseurs de ressources linguistiques, provenant des sociétés industrielles comme du secteur public de recherche, discuteront des priorités et des aspects économiques pour la production, la distribution et l'utilisation de ressources linguistiques, ainsi que de l'importance de leur mise à disposition.

• *Présidence :* Khalid Choukri (ELRA). • *Participants :* David Brooks (Microsoft) , J.P. Chanod (Xerox), Claudio Cirilli (Synthema), Melvyn Hunt (Dragon Systems), Ian Johnson (Sharp), Siegfried Kunzmann (IBM-Europe), Nils Lenke (PHILIPS), Jan Odijk (Lernout & Hauspie Speech products).

# Programme synthétique

## Jeudi 28 mai 1998

**10h00 – 11h40** — **Session inaugurale**

**11h40 – 12h00** — **Pause**

**12h00 – 13h20** — **4 Sessions en parallèle**

*Session A : Ressources linguistiques*
*Session B : Evaluation en traduction automatique*
*Session C : Table ronde "Maintenance des RL"*
*Session D : Evaluation des systèmes de dialogues oraux*

**13h20 – 14h40** — **Déjeuner**

**14h40 – 16h40** — **4 Sessions en parallèle**

*Session E : Acquisition de lexiques*
*Session F : Evaluation en TLN*
*Session G : RL : Considérations politiques*
*Session H : Evaluation des systèmes de dialogues oraux*

**16h40 – 17h00** — **Pause**

**17h00 – 18h20** — **Table ronde 1**

*Table ronde des agences de financement*

**18h20 – 18h30** — **Pause 10 min.**

**18h30 – 19h30** — **Table ronde 2**

*Table ronde sur la coopération entre l'UE et les autres pays dans le domaine des RL et de l'évaluation*

## Vendredi 29 mai 1998

**9h00 – 9h40** — **2 Orateurs invités en parallèle (1), (2)**

**9h40 – 10h40** — **4 Sessions en parallèle**

*Session I : Projets lexicaux*
*Session J : Evaluation d'outils et outils d'évaluation*
*Session K : Traitement de la parole et évaluation*
*Session L : Projets de RL orales*

**10h40 – 11h00** — **Pause**

**11h00 – 12h00** — **Sessions I, J, K, L (suite)**

**12h00 – 13h20** — **4 Sessions affichées en parallèle**

*P1 : Lexique*
*P2 : Evaluation*
*P3 : Corpus*
*P4 : Bases orales et lexiques phonétiques*

**13h20 – 14h40** — **Déjeuner**

**14h40 – 16h40** — **4 Sessions en parallèle**

*Session M : Projets lexicaux : réseaux sémantiques*
*Session N : Evaluation : lemmatiseurs et analyseurs syntaxiques*
*Session O : Projets de corpus*
*Sessions P & Q : Projets de RL orales • RL : aspects stratégiques*

**16h40 – 17h00** — **Pause**

**17h00 – 18h40** — **2 Tables rondes en parallèle**

*Utilisation industrielle et en R&D des RL*

*Table ronde EAGLES sur les standards lexico-sémantiques pour les systèmes d'information*

(1) *Ressources orales et écrites en Europe*, Nicoletta Calzolari et Harald Höge.
(2) *Problèmes d'évaluation des systèmes en langage naturel*, Margaret King et Bente Maegaard.

## Samedi 30 mai 1998

**9h00 – 9h40** — **2 Orateurs invités en parallèle (1), (2)**

**9h40 – 10h40** — **4 Sessions en parallèle**

*Session R : Ontologies & bases de connaissances*
*Session S : Evaluation : tâches et composantes*
*Session T : Outils du TLN*
*Session U : Evaluation de systèmes vocaux*

**10h40 – 11h00** — **Pause**

**11h00 – 12h00** — **Sessions S, T, U (suite)**

**12h00 – 13h20** — **4 Sessions affichées en parallèle**

*P5 : Lexiques, corpus et terminologie*
*P6 : Evaluation*
*P7 : Applications*
*P8 : Outils et formats pour les RL*

**13h20 – 14h40** — **Déjeuner**

**14h40 – 16h40** — **4 Sessions en parallèle**

*Session W : Terminologie*
*Session X : Graphes lexicaux*
*Session Y : Aspects multilingues*
*Session Z : Evaluation des systèmes vocaux*

**16h40 – 17h00** — **Pause**

**17h00 – 18h30** — **Clôture**

(1) *Le projet TREC et l'aspect multilingue,* Donna Harman et Gregory Grefenstette.
(2) *Problèmes de dialogue Homme-Machine*, Christian Dugast (un second orateur sera annoncé prochainement).

# Vérification de textes en anglais simplifié de l'AECMA et en langage contrôlé, *Richard H. Wojcik, Boeing*


## Résumé

L'anglais simplifié de l'AECMA est un des exemples les plus connus de langage contrôlé, mais il pose des problèmes particuliers lorsqu'il s'agit de mettre au point un programme de vérification de langage contrôlé capable de le traiter. En dépit de ces difficultés, un vérificateur pour l'anglais simplifié de l'AECMA est un outil extrêmement utile pour aider les auteurs de textes à se conformer à la norme de ce langage. Cet article présente la norme pour l'anglais simplifié de l'AECMA et s'intéresse plus particulièrement aux programmes de vérification, notamment au Vérificateur d'Anglais Simplifié développé à Boeing.


## Introduction

Ces dernières années, le secteur industriel a vu se développer des normes pour la rédaction de documentations techniques, connues sous le nom de langages contrôlés. Pour simplifier, les langages contrôlés sont des normes de rédaction pour lesquelles la grammaire, le style et le vocabulaire sont plus restreints que dans le cas de la rédaction de textes techniques normaux. Typiquement, ils sont basés sur un vocabulaire de moins de 1 000 mots, auxquels les fabricants rajoutent leur propre vocabulaire technique avec certaines contraintes. Par exemple, l'anglais simplifié de l'AECMA stipule 20 catégories de termes techniques possibles et 6 catégories de procédés de fabrication. Les contraintes imposés sur le style et la grammaire tendent à forcer les auteurs à rédiger sous forme de phrases et de paragraphes plus courts et plus clairs. Il est également fréquent que les langages contrôlés imposent l'utilisation d'articles et interdisent le recours à des formes nominales ou verbales complexes.

La tendance actuelle au développement de normes pour la rédaction en langage contrôlé puise son origine dans le fameux Basic English (Anglais de Base) dont le concept a été proposé dans les années 30 par K.C. Ogden, dans le but d'aider à l'apprentissage de l'anglais. Cependant, les industriels s'intéressent plus aux langages contrôlés comme un moyen de produire des manuels d'entretien et de réparation qui peuvent être plus faciles à comprendre par des non anglophones ou plus faciles à traduire. Pour leur part, les spécialistes en linguistique automatique perçoivent les langages contrôlés comme étant à la fois un objet d'étude et un débouché pour leurs travaux. Les langages contrôlés sont conçus pour satisfaire les besoins du lecteur, mais ils ne sont pas particulièrement commodes pour celui qui rédige le texte. Aussi, les entreprises qui ont adopté des langages contrôlés ont-elles besoin de programmes de vérification automatique de textes écrits dans ces langages, afin de réduire les coûts de formation tout en améliorant la cohérence et la précision de la rédaction. De même, un texte écrit dans un style cohérent et avec une syntaxe prédictible présente un intérêt certain lors du processus de traduction, surtout s'il est automatique.

L'Anglais Simplifié de l'AECMA est probablement l'exemple de langage contrôlé le plus connu. En effet, c'est un standard à l'échelle industrielle. Les industries aérospatiales ont l'obligation contractuelle de s'y conformer. En 1979, l'AEA (Association des Compagnies Aériennes Européennes – *Association of European Airlines*) a demandé à l'AECMA (*European Association of Aerospace Manufacturers*) d'évaluer la lisibilité de ses documentations. L'AECMA a mis au point l'Anglais Simplifié à partir d'exemples de normes antérieures de langages contrôlés, de guides pour la rédaction de documents dans le domaine aérospatial et de normes d'écritures du secteur militaire. La première version de la norme (document AECMA PSC-85-16598) a été publiée en 1986 et l'ATA (Association pour le Transport Aérien) d'Amérique l'a rendue obligatoire en tant que spécification pour l'édition de documents (Spécification ATA n° 100). Plus récemment, la spécification connue sous le nom AECMA 1000D a également rendu obligatoire l'utilisation de l'Anglais Simplifié pour les équipements militaires du domaine de l'aérospatiale vendus dans la Communauté Européenne. La version actuelle du Guide de l'Anglais Simplifié, publiée en janvier 1998, est dénommée *Issue 1, Rev 1*. Pour toute information sur la façon de se procurer le Guide de l'Anglais Simplifié, consulter le site Web de l'AECMA (http://www.aecma.org).

## La norme de l'Anglais Simplifié

Comme nous l'avons évoqué ci-dessus, le vocabulaire de base de l'Anglais Simplifié contient moins de 1 000 mots. Seuls 200 d'entre eux sont des verbes. Les fabricants peuvent ajouter des « Termes Techniques » et des « Procédés de Fabrication » à ce vocabulaire (environ 7 000 termes pour l'instant, dans le cas de Boeing). La façon dont l'Anglais Simplifié est conçu est telle qu'il est relativement facile de rajouter des noms ou des adjectifs mais qu'il est difficile de rajouter des verbes. Ainsi, l'auteur d'un texte est par exemple contraint d'écrire « *Do the Leak Test* » (Faire le Test de Fuites) plutôt que d'utiliser la tournure moins explicite « *Test for leaks* » (Tester les fuites). La référence à un « *Leak Test* » (Test de Fuites) sous-entend qu'il existe un ensemble précis d'opérations à effectuer pour vérifier l'absence de fuites.

L'utilisation extensive de structures nominales en Anglais Simplifié de l'AECMA est tout à fait intentionnelle. La norme a été principalement conçue pour des textes spécifiant des procédures et non pour des textes descriptifs. La norme ne rend certes pas justice à la richesse des structures verbales, ce qui complique la tâche aux auteurs qui veulent exprimer des relations temporelles complexes dans une seule et même phrase. Au contraire, l'auteur est contraint de faire appel à des adverbes ou à des prépositions plus explicites, telles que *before* (avant) ou *after* (après) pour exprimer ces relations temporelles. En ce qui concerne les conjugaisons, les temps simples sont généralement suffisants pour spécifier des procédures, alors que les formes verbales plus complexes sont plus utiles lorsqu'il s'agit de décrire le fonctionnement d'un système ou les flux d'un processus.

Le paragraphe 6 du chapitre 1 du Guide de l'Anglais Simplifié impose de fait quelques restrictions aux textes descriptifs, mais l'accent est surtout mis sur le relâchement des règles qui s'appliquent au langage pour spécifier les procédures. Par exemple, l'interdiction des tournures passives est quelque peu assouplie et les auteurs ont droit à 5 mots de plus dans leurs phrases, c'est-à-dire une longueur maximale portée à 25 mots. En fait, toutes les industries aérospatiales n'appliquent pas, dans leurs manuels d'entretien, l'Anglais Simplifié de l'AECMA de façon identique pour les textes descriptifs. La tendance actuelle consiste à appliquer l'Anglais Simplifié à tous les types de textes dans les manuels d'entretien et dans les documents associés, mais certaines compagnies ont limité l'utilisation de la norme aux procédures d'entretien proprement dites et aux notes descriptives qui les accompagnent.

Il existe environ 55 règles d'écriture dans le Guide de l'Anglais Simplifié. Celles qui suivent sont parmi les plus connues :

- La longueur maximale d'une phrase est de 20 mots (25 pour les textes descriptifs)
- La longueur maximale d'un paragraphe est de 6 phrases
- La longueur maximale d'un nom composé est de 3 mots
- L'utilisation de *be* dans des formes verbales progressives et de *have* dans des formes verbales perfectives est interdite
- L'utilisation de la voix passive est interdite dans la spécification de procédures et déconseillée dans les textes descriptifs
- L'usage du gérondif (forme en *-ing*) est interdit
- Les étapes distinctes d'une procédure doivent être spécifiées dans des phrases distinctes
- Les mots ne peuvent être utilisés que dans leur acception prévue
- Les articles tels que *a*, *an* et *the* doivent être utilisés aussi souvent que possible

Nombre de restrictions imposées par l'Anglais Simplifié ressemblent à celles que l'on peut trouver dans les manuels de recom-

mandations pour la rédaction de textes techniques. Il y a des règles qui requièrent la présence de phrases focales dans chaque paragraphe (règle 6.5) et de mots de liaison entre les phrases (règle 6.6). Une des règles (la règle 6.8) recommande même que les informations nouvelles et complexes soient introduites lentement et progressivement. Il faut préciser que ce langage contrôlé particulier n'a pas été conçu en prenant en compte l'éventualité de son utilisation dans le contexte d'applications informatiques. Les concepteurs et les auteurs de ce standard se sont concentrés sur les aspects propres à la rédaction de manuels d'entretien des avions, problème qui pose des difficultés spécifiques pour les lecteurs non anglophones.

La norme de l'Anglais Simplifié est régulièrement examinée et remise à jour par les associations de constructeurs du secteur aérospatial en Europe et en Amérique du Nord. Le groupe SEMG (Simplified English Maintenance Group) de L'AECMA se réunit une ou deux fois par an pour discuter d'éventuelles modifications et pour tenter d'anticiper les besoins des industriels.

## Programmes de vérification automatique de l'Anglais Simplifié

Les règles d'écriture de l'Anglais Simplifié de l'AECMA ne sont pas difficiles à comprendre. Une formation de base sur ce langage prend typiquement 2 ou 3 jours, mais il faut compter au moins 1 mois avant qu'un auteur ne maîtrise la rédaction de procédures et 3 mois pour les textes descriptifs. Les applications informatiques peuvent accélérer le processus d'apprentissage en apportant aux auteurs une indication sur la conformité de leur texte pour les aspects les plus mécaniques de la norme (par exemple, la longueur des mots, des phrases ou des paragraphes, la ponctuation, la structure grammaticale). Il est certes très difficile d'écrire un programme qui détecterait des différences subtiles dans l'utilisation des mots, mais un bon analyseur syntaxique est capable de repérer des problèmes grammaticaux tels que l'utilisation d'un mot dans une mauvaise classe syntaxique.

La question de l'intérêt d'une automatisation complète du processus de vérification reste un sujet controversé. Sans nul doute, les auteurs souhaiteraient que la tâche de rédaction soit aussi automatisée que possible. Cependant, il est facile de sous-estimer la difficulté à réécrire des textes en Anglais Simplifié. Un auteur peut remplacer un mot interdit par une des alternatives proposées dans le Guide de l'Anglais Simplifié, mais c'est bien plus difficile pour un programme informatique que pour un être humain de choisir la meilleure des alternatives proposées. Pire, les alternatives du Guide contiennent souvent des mots qui appartiennent à une classe grammaticale différente de celle du mot d'origine. La sélection d'une alternative particulière peut donc entraîner le remaniement complet de la structure de la phrase. En conséquence, nous avons adopté le point de vue de fournir à l'utilisateur des informations qui l'aident dans le contexte d'un processus de révision qui demeure manuel.

## Le Vérificateur d'Anglais Simplifié de Boeing

Après la mise en place de la première version de la norme de l'Anglais Simplifié, en 1986, Boeing a décidé de réécrire ses manuels d'entretien pour les avions des modèles 737, 747, 757 et 767. Comme il n'existait pas de vérificateurs d'Anglais Simplifié dans le commerce, Boeing a développé le sien, en 1989. Le vérificateur de Boeing fonctionne en mode interactif ou en batch (non interactif) sur un grand nombre de plates-formes informatiques. Il prend un texte en entrée et génère en sortie des diagnostics d'erreurs. Voici un exemple de rapport produit par le vérificateur :

*The instruments listed below were used to prepare this procedure* (Les outils listés ci-dessous ont été utilisés pour préparer cette procédure)

Erreur voix passive :
were used
Erreur d'adverbe :
Below
Autorisé comme : préposition
Erreur de verbe :
Listed
Utiliser : record
list (n)

Le rapport ci-dessus indique à l'auteur qu'il doit reconstruire sa phrase avec utilisation de la voix active. De plus, il révèle que le mot *below* est autorisé comme préposition, mais que ce n'est pas ainsi qu'il est utilisé dans la phrase. Enfin, on remarque qu'il existe 2 alternatives acceptables pour le mot *listed*. Une façon de rendre conforme la phrase de départ consiste à écrire : *We used the instruments that follow to prepare this procedure* (Nous avons utilisé les outils qui suivent pour préparer cette procédure). Il existe certainement d'autres façons de remanier cette phrase, mais notons que la nouvelle formulation ne contient aucun mot correspondant au verbe *list*.

L'analyseur synthaxique sur lequel repose le vérificateur automatique de Boeing a été développé en interne. Il est basé sur une grammaire de type GPSG contenant plus de 350 règles et il produit le graphe de toutes les hypothèses possibles. La meilleure hypothèse est sélectionnée par un système de pondérations numériques associées aux règles et aux entrées lexicales. Dans la mesure où l'objectif du vérificateur automatique consiste moins à identifier les erreurs de grammaire que d'identifier dans les phrases correctes une non conformité à la norme, la détection s'opère simplement en indiquant les structures et les mots sur lesquels il faut attirer l'attention de l'utilisateur. La version du système actuellement utilisée ne fait que marginalement appel à l'information sémantique, mais un système expérimental récemment développé permet de désambiguïser les graphes d'hypothèses sur la base des pondérations numériques et du sens des mots. Ce système n'a pas été utilisé pour « traduire » un texte en Anglais Simplifié, mais pour signaler l'utilisation impropre de certains mots et pour proposer un diagnostic plus intelligent sur la façon de corriger le texte.

## Conclusions

L'intérêt d'un vérificateur automatique d'Anglais Simplifié ne réside pas tant dans sa capacité à produire une analyse exhaustive de la conformité à la norme ; c'est plutôt dans l'aide qu'il fournit à l'utilisateur dans l'accomplissement de sa tâche de rédaction de textes de qualité en Anglais Simplifié. Non seulement la correction automatique est difficile à effectuer, mais on peut même s'interroger sur son intérêt. Ce qui est utile, ce sont des programmes qui permettent aux auteurs d'apprendre plus vite à manier la norme et qui les aident à maintenir une qualité de rédaction constante, même sous la pression de délais parfois très courts. C'est surtout dans le contexte de la réécriture en Anglais Simplifié de gros volumes de textes déjà existants, que l'utilité des vérificateurs automatiques est la plus flagrante.

Les langages contrôlés sont souvent associés à la traduction automatique. L'industrie aérospatiale n'a pas, pour l'instant, prêté beaucoup d'attention à ce domaine connexe. Historiquement, l'anglais a toujours été la langue dominante dans le secteur industriel et les clients ont l'habitude de prendre à leur charge les coûts de traduction. C'est pourquoi, très peu d'efforts ont été faits pour adapter un système de traduction automatique à l'Anglais Simplifié de l'AECMA. En fait, certaines compagnies aériennes ont le sentiment que l'Anglais Simplifié permet de produire des manuels d'entretien dans lesquels l'anglais était suffisamment simple pour que des mécaniciens expérimentés puissent les utiliser sans avoir recours à une traduction. Il demeure cependant une forte demande pour des systèmes d'aide à la traduction et cette demande va probablement croître au fur et à mesure des améliorations techniques que vont connaître les systèmes de traduction automatique. D'après nous, les vérificateurs automatiques d'Anglais Simplifié sont des précurseurs indispensables pour une traduction automatique efficace. La qualité de l'Anglais Simplifié doit être bonne et homogène pour que les systèmes de traduction automatique puissent fonctionner correctement.

Richard H. Wojcik
PO Box 3707,
MS 7L-43
Seattle, WA 98006,
USA
richard.h.wojcik@boeing.com

# Le résumé automatique de texte et son évaluation

*Frances C. Johnson, Manchester Metropolitan University*

Résumé

Cet article présente un bref panorama des recherches en résumé automatique de texte (et sur les domaines très voisins de l'extraction de texte et de la contraction automatique de texte)[1]. Nous mettons l'accent sur les problèmes relatifs à l'évaluation, un sujet d'actualité dont témoigne l'annonce de la conférence SUMMAC (SUMMArization Conference). Cette manifestation est organisée dans le cadre du programme DARPA TIPSTER (Text Program) qui vise à promouvoir l'évaluation de ces technologies selon une approche tournée vers l'utilisateur. Les problématiques évoquées dans cet article sont discutées en détail dans des publications plus spécifiques[2] auxquelles le lecteur est invité à se référer. L'évaluation est incontestablement une question à l'ordre du jour qui conditionne le développement de systèmes de plus en plus performants. Ce domaine de recherche devrait donc bénéficier pleinement de la mise en place de mesures de performances bien définies.

L'objectif des techniques de résumé et de contraction automatique de texte (au sens le plus large) consiste à produire une représentation concise d'un texte qui en préserve le message principal. Cette tâche est loin d'être triviale et elle requiert la mise en œuvre de techniques de compréhension et de génération de texte. Pour relever ce défi, des chercheurs de différents domaines ont joint leurs efforts. Les premiers travaux sur le sujet remontent à la fin des années 1950, lorsque des chercheurs en Sciences de l'Information ont tenté de déterminer dans quelle mesure les techniques utilisées pour l'indexation de documents pouvaient s'appliquer à la tâche de génération automatique de résumé de texte. En s'appuyant sur l'extraction de phrases-clefs, l'approche en question fait appel à des méthodes statistiques et/ou de reconnaissance des formes pour identifier, dans les textes, les phrases porteuses de sens qui servent ensuite de base à la génération automatique du résumé. A un niveau de traitement plus élaboré, des techniques d'analyse grammaticale (partielle) peuvent être employées pour traiter les problèmes de cohérence et de cohésion du texte, notamment pour identifier les relations rhétoriques, les progressions thématiques et pour résoudre ou éliminer les expressions référantes[3]. En général, ces méthodes de résumé automatique s'avèrent robustes au sens où elles fonctionnent avec des textes de domaines quelconques. Cependant, l'intelligibilité du résumé produit n'est pas toujours garantie. Un second type d'approches est basé sur les techniques d'intelligence artificielle (et les disciplines voisines) et fait appel à des méthodes automatiques de compréhension du discours et de génération de textes écrits. Typiquement, ces approches tentent de produire des phrases qui résument le contenu d'un texte en utilisant différentes représentations des connaissances sur le domaine, par exemple des réseaux sémantiques[4]. En tant qu'application des techniques de compréhension de langage naturel, cette tâche permet d'illustrer les capacités du système à comprendre un texte en s'appuyant sur un modèle des processus de compréhension humaine.

L'évaluation des techniques de résumé et de contraction automatique de texte, au sens de la mesure des performances du système par l'intermédiaire d'un score, est une problématique complexe. Il est néanmoins intéressant de noter que ce sont les techniques à base d'extraction de phrases-clefs qui prévalent dans les systèmes commerciaux récents, notamment les systèmes ConText d'Oracle, Searchable Lead de Mead Data Central et Text Summarizer de British Telecom. En outre, des évaluations précédentes ont mis en évidence que de simples heuristiques consistant à extraire localement certains passages du texte, fournissent la plupart du temps des résultats optimaux (voir par exemple Kupiec 1995). L'objectif central de ces systèmes (qui guide donc le processus d'évaluation) consiste à identifier les informations les plus importantes et les plus pertinentes dans un texte. Cet objectif est commun aux techniques de résumé automatique, de contraction automatique et d'extraction d'information. Il n'est donc pas très surprenant d'observer des similitudes importantes entre les techniques utilisées. Cependant, des distinctions subtiles doivent être effectuées et celles-ci affectent certains détails de l'approche adoptée en évaluation. L'extraction d'information se limite à analyser en profondeur les parties d'un document qui sont susceptibles de contenir des informations qui se rattachent à un ensemble de catégories précises et de sujets particuliers, fixés au préalable. Dans ce cas, les mesures de performances comparatives sont assez solidement établies et elles sont basées sur l'évaluation des capacités du système à retrouver l'information pertinente, en termes de rappel et de précision. Le taux de rappel correspond à la proportion de données pertinentes effectivement extraites alors que le taux de précision s'évalue comme la fraction des informations pertinentes parmi celles qui ont été extraites. Ces mesures (ou d'autres, similaires, qui permettent de rendre compte de la sur-génération ou de l'omission d'information) peuvent être utilisées pour l'évaluation des méthodes de résumé automatique de texte, dans le cadre desquelles les mesures pour l'évaluation sont calculées à partir de la comparaison ou de la concordance des phrases sélectionnées avec une représentation de référence ou un ensemble de phrases résumant le texte. Cependant, en l'absence de critères d'extraction bien définis, il est délicat de déterminer les informations qui sont pertinentes et celles qui ne le sont pas. Ceci est encore plus vrai dans le cas de la contraction automatique de texte, car il existe de nombreuses façons de représenter validement le contenu d'un document, chacune correspondant à un sous-ensemble de phrases différentes. De plus, le degré de pertinence de certaines informations varie selon les individus, notamment en fonction de leurs besoins et de leurs centres d'intérêts. C'est pourquoi il apparaît que cet effort de calibration apporte finalement peu d'information sur le caractère fonctionnel et sur l'utilité des résumés produits. En d'autres termes, il est difficile de mesurer de cette façon si un résumé reflète réellement le message contenu dans le texte d'origine et s'il fournit une réponse appropriée et une aide utile aux lecteurs dans le cadre de l'accomplissement d'une tâche donnée. Les résumés de textes constituent des outils essentiels pour la recherche et l'accès aux ressources textuelles : ils sont particulièrement utiles à l'issue d'une recherche documentaire qui retourne une liste de documents potentiellement intéressants pour l'utilisateur, qui doit alors sélectionner ceux qui sont le plus appropriés. Compte tenu du développement actuel des technologies de la communication, lesquelles rendent de plus en plus accessibles des volumes croissants d'informations diverses, les outils permettant d'assister l'utilisateur dans la recherche et le filtrage d'informations spécifiques connaissent eux aussi un essor spectaculaire. L'aspect utilisateur ne peut donc plus être négligé dans le processus d'évaluation.

Pour aller dans le sens d'une évaluation des fonctionnalités des systèmes de résumé automatique, il convient de définir à un autre niveau la notion de performance, afin de mieux comprendre quelles sont les caractéristiques essentielles qui rendent un résumé utile, c'est-à-dire qui accroissent les chances que celui-ci remplisse effectivement les fonctions que l'on en attend. D'une part, les performances de la recherche d'information sont conditionnées par la qualité intrinsèque des résumés : le critère de qualité est alors lié à l'identification de textes pertinents par rapport à une requête particulière. D'autre part, les performances de recherche des utilisateurs est aussi un facteur d'évaluation des performances : il convient alors de les évaluer en terme d'aide apportée à l'utilisateur pour lui permettre d'effectuer des jugements sur la pertinence d'un texte par rapport à ses besoins. Dans ce cas, il est nécessaire de mesurer le caractère informatif d'un résumé pour un utilisateur particulier et une tâche précise, ainsi que d'évaluer l'impact d'autres caractéristiques, telles que la présentation du résultat.

La conférence SUMMAC, soutenue par DARPA TIPSTER (Text Program), permettra aux chercheurs de réfléchir à la définition et à l'amélioration de métriques pour l'évaluation et la comparaison des performances de systèmes de résumé automatique de texte. L'approche proposée ne présuppose pas l'exis-

tence d'un seul résumé correct mais est plutôt basée d'une part sur une mesure du temps nécessaire à l'utilisateur pour juger de la pertinence du résumé (le résumé capture-t-il l'information recherchée par un utilisateur, telle qu'elle a été exprimée lors d'une requête ?) et d'autre part sur des décisions de catégorisation (les concepts-clefs du texte sont-ils présents dans le résumé produit ?). Une des mesures de la valeur informative du substitut du texte s'évalue donc comme la proportion des jugements de pertinence qui sont identiques à ceux effectués en ayant accès au texte entier. Des évaluations quantitatives supplémentaires, basées sur des jugements de préférence émis par l'utilisateur, seront également effectuées selon un ensemble de critères d'acceptabilité. Cette approche constitue un développement tout à fait heureux vers une évaluation centrée sur l'utilisateur. Il est cependant indispensable que les contributions individuelles au développement de métriques d'évaluation soient poursuivies. En particulier, un programme de recherche sera proposé pour étudier quelles sont les caractéristiques des résumés qui induisent les différents jugements observés. Ce programme cherchera également à déterminer quelles sont les conditions auxquelles un système atteint des performances jugées optimales, et pourquoi il en est ainsi. Ce n'est qu'à ce moment-là que l'intérêt des techniques de résumé et de contraction automatique de texte sera réellement perçu et que les utilisateurs pourront, raisonnablement, tirer parti d'une présentation des informations sous forme concise.

### Références bibliographiques

*1. F.C. Johnson, C.D. Paice, W.J. Black, A.P. Neal : The application of linguistic processing to automatic abstract generation. Journal of Document and Text Management, 1993, vol. 1 (3), pp. 215-242. (Ré-imprimé dans : Karen Spark Jones and Peter Willett (Eds). Readings in Information Retrieval. San Francisco. Morgan Kauffman Publishers, 1997, pp. 538-553.)*

*2. Julian Kupiec, Jan Pedersen, Francine Chen : A trainable document summarizer. Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, juillet 1995, pp. 68-73.*

*3. C.D. Paice : Constructing literature abstracts by computers : techniques and prospects. Information Processing and Management, 1990, vol. 26 (1), pp. 171-186.*

*4. Summarizing text for intelligent communication. Seminar Program, décembre 1993.*

*5. Intelligent scalable text summarization. Proceedings of a workshop sponsored by the Association of Computational Linguistics. Madrid (Espagne), juillet 1997.*

*6. DARPA TIPSTER Text Program. http://www.tipster.org*

Dr. Frances C. Johnson
Departement Information & Communications
Manchester Metropolitan University
Geoffrey Manton Building
Rosamond Street West
Manchester M15 6LL, Royaume Uni
Courriel : F.Johnson@mmu.ac.uk

Notes :

1. Bien qu'il existe des nuances entre les notions de résumé, de contraction et d'extraction de texte, celles-ci ne sont pas explicitées dans cet article ; nous nous limitons à une présentation d'ensemble des techniques utilisées pour identifier les thèmes principaux d'un texte

2. F.C. Johnson : « A critical review of system-centred to user-centred evaluation of automatic abstracting techniques ». Article en préparation

3. Pour un panorama complet de ces techniques, voir Paice 1990, Johnson et al. 1993 et Kupiec 1995

4. Plusieurs techniques et applications de ce type sont présentées dans des actes de workshops sur le résumé automatique de texte (1993 [4] et 1997 [5])

# MultiMeteo : production de bulletins météorologiques multilingues

*José Coch, ERLI*

### Objectifs

Le but du projet MultiMeteo est de développer un système de génération automatique de comptes rendus multilingues et d'utiliser ce système pour produire des bulletins météorologiques.

Ce système vise à rendre les bulletins météorologiques plus accessibles aux utilisateurs et plus adaptés à leurs besoins, du fait notamment qu'ils puissent être dans la langue de leur choix (pour le moment, en allemand, anglais, espagnol, français et néerlandais).

MultiMeteo est un projet sur 3 ans financé en partie par le programme Ingénierie Linguistique (LE) de la Commission Européenne et en partie par des agences météorologiques européennes. Le projet implique une quarantaine de personnes en France, en Espagne, en Autriche et en Belgique. On dénombre environ 15 sites pilotes en Europe, qui présentent différentes caractéristiques climatiques (sud, nord, plaines, montagne, mer, etc...). Dans chacun des sites, plusieurs styles de bulletins sont pris en compte (local ou régional, pour le grand public ou pour une audience particulière (par exemple les sportifs), bord de mer, etc...).

### Motivations

Le marché des bulletins météorologiques en Europe ne se limite plus au niveau national. Le volume des déplacements croît rapidement qu'il s'agisse de voyages d'agrément ou professionnels.

En outre, avec le développement du multimédia et la libéralisation du secteur des télécommunications, le marché des bulletins météorologiques déborde de plus en plus les frontières des états. Les agences météorologiques doivent distribuer leurs bulletins en plusieurs langues et ont donc besoin d'outils multilingues.

Les volumes traités sont très élevés et les délais de production d'informations fraîches et utiles sont de plus en plus courts : des centaines de bulletins sont produits plusieurs fois par jour par chaque agence météorologique.

Les utilisateurs finaux deviennent eux aussi de plus en plus exigeants en termes de qualité et d'adéquation à leurs besoins. Dans ce contexte, l'adaptation aux besoins est réalisée en produisant plusieurs styles de bulletins, ceux-ci étant générés à partir des mêmes données.

### Situation du marché

En Europe, le marché de l'information météorologique n'existe que depuis une vingtaine d'années, alors que cela fait 50 ans qu'il se développe aux Etats-Unis. Néanmoins, le « bulletin météo » est une des émissions les plus populaires à la télévision et il constitue même le thème principal de certaines chaînes de télévision par satellite ou par câble. Trois facteurs sont à l'origine de la croissance récente que l'on observe :

- Précision accrue des informations en raison de progrès techniques en physique, en informatique, en technologie des satellites et des radars, etc.
- Développement de la télévision, des télécommunications et maintenant du multimédia,
- Réduction des services financés par l'état, ce qui favorise l'émergence de fournisseurs de services dans le secteur privé.

Il convient de distinguer 3 marchés différents : le grand public, les professionnels et les médias.

Le marché du grand public est composé des consommateurs directs des bulletins météorologiques, ceux-ci étant distribués par l'intermédiaire de réseaux et de terminaux bon marchés, facilement accessibles et moyennant le paiement des prestations. Audiotex bénéficie d'une audience importante en Europe, alors que Videotex connaît un succès comparable en France, grâce au Minitel. Internet a plus d'impact que videotex dans le reste de l'Europe.

Pour le grand public, le style et la présentation de l'information jouent un rôle essentiel et les bulletins doivent être adaptés à l'endroit qu'ils couvrent, notamment s'il s'agit de stations balnéaires ou de sports d'hiver.

Le marché des professionnels peut être décomposé en plusieurs sous-marchés, selon le secteur d'activités considéré : agriculture, transports, commerce, bâtiment, tourisme, etc. Chaque secteur requiert des données

météorologiques bien précises telles que la vitesse du vent, la température à des endroits non-standards, le degré d'humidité ou le volume des précipitations et leur localisation.

Il est donc primordial de prendre en compte des phénomènes divers et variés.

Le marché des médias s'adresse, au bout du compte, au grand public, sauf dans quelques cas où ce sont les professionnels qui sont ciblés. La compétition est intense, y compris de la part de fournisseurs de services hors Europe et la demande pour des produits « clés en main » va croissant. Avec l'extension du câble et de l'Internet, les bulletins météorologiques locaux sont de plus en plus recherchés.

## Situation actuelle

Les bulletins météorologiques sont produits en s'appuyant sur différents niveaux d'expertise humaine qui vont de simples paramètres physiques ou climatiques (pression, vitesse et direction du vent, température, humidité, risques de pluie, neige ou grêle, couverture nuageuse, etc...) calculés par des modèles mathématiques sophistiqués tournant sur des super-ordinateurs, jusqu'à la description littérale du temps, à l'intention des utilisateurs finaux. De telles informations font donc appel à la fois à des technologies de pointe et à l'expérience humaine.

Deux procédés sont utilisés pour produire les prévisions :

- L'application de lois de transformations physiques à un état initial de l'atmosphère. Cette approche repose essentiellement sur un ensemble de mesures ponctuelles simultanées à partir desquelles l'application d'un modèle mathématique permet de calculer, pas à pas, les états les plus probables dans le futur.
- Le choix entre le résultat de plusieurs simulations (provenant par exemple de modèles différents) ou la prise en compte de connaissances sur les conditions locales (observations, orographie ou effets dus à la proximité d'un centre urbain).

Des fichiers de données et des diagrammes sont mis à la disposition de la presse au moins deux fois par jour, avec en général des prévisions à 5 jours. Les centres de météorologie produisent des bulletins locaux plusieurs fois par jour pour l'Audiotex, le Videotex et l'Internet. Cependant, en raison des volumes que cela représente, ceux-ci ne sont généralement pas traduits dans des langues étrangères.

## Spécifications techniques

Les données en entrée sont constituées d'un ensemble de tableaux (un par zone géographique) dont les lignes représentent 3, 6 ou 12 heures et dont les colonnes correspondent aux paramètres météorologiques tels que l'humidité, la température, la couverture nuageuse, la vitesse du vent et sa direction. Les données en entrée sont fournies par les ordinateurs de l'Institut Météorologique.

Les données en sortie sont des fichiers aux formats HTML ou RTF qui contiennent un ou plusieurs bulletins météorologiques, dans un format lisible par les logiciels de traitement de textes et les outils de navigation sur le Web.

Les mêmes données en entrée peuvent être utilisées pour générer plusieurs bulletins météorologiques dans différents styles et en différentes langues.

D'un point de vue du matériel et du système d'exploitation, ce sont des PC sous Windows 95 ou Windows NT et des stations Unix sous Solaris 2 qui sont utilisés.

## Contexte

Le consortium est composé de 2 types de partenaires : les utilisateurs et les fournisseurs.

Les utilisateurs sont constitués de 4 services météorologiques nationaux en Europe : Météo France, l'Instituto Nacional de Meteorología (Espagne), l'Institut Météorologique Royal (Belgique) et le Zentralanstalt für Meteorologie und Geodynamik (Autriche). Aucun de ces instituts ne se trouve en situation de monopole, mais tous jouissent d'une position dominante au niveau national en ce qui concerne la production d'informations météorologiques.

Les fournisseurs sont ERLI, une société européenne en pointe dans le secteur de l'Ingénierie Linguistique, et CL Servicios Lingüísticos, une autre société au statut similaire dans le domaine de la Terminologie.

Le noyau du système de génération de bulletins est basé sur la toolbox AlethGen d'ERLI qui a déjà été utilisée pour produire des textes dans des applications réelles, notamment des réponses automatiques à des réclamations pour une société française de vente par correspondance (voir [Coch, David et Magnole 95]). AlethGen a été développé entre 1993 et 1995.

Les principales caractéristiques d'AlethGen sont les suivantes :

- La qualité des textes générés : les textes produits sont d'excellente qualité, en termes de fluidité, de clarté et d'adéquation aux besoins (voir [Coch 1996b]).
- La modularité : il s'agit d'un logiciel industriel qui utilise plusieurs techniques de façon hybride ; il est constitué de plusieurs modules qui peuvent être intégrés et utilisés de différentes façons pour satisfaire diverses spécifications liées à l'application (voir [Coch 1996a]).

AlethGen est composé d'un moteur (écrit en C++) et de bases de connaissances. L'approche s'inspire essentiellement d'une théorie exposée par Mel'cuk dans [Mel'cuk 88].

Les entrées terminologiques en allemand, anglais, espagnol, français et néerlandais, sont fournies par CL Servicios Lingüisticos.

## Architecture

Le logiciel MultiMeteo est composé de 2 modules principaux (appelés stations) :

- La station de Production de Texte
- La station d'Administration

La station de Production de Texte est basée sur la toolbox AlethGen d'ERLI. Comme nous l'avons indiqué plus haut, AlethGen est composé d'un moteur (écrit en C++) et de bases de connaissances. Un premier sous-module lit les entrées numériques, calcule la structure et le contenu des textes à générer (dans le style sélectionné) et produit une formule conceptuelle. Ce premier sous-module s'appelle le Planificateur. Un second module prend en entrée la formule conceptuelle et produit les textes du bulletin météorologique. Ce second sous-module s'appelle le Réalisateur. D'un point de vue linguistique, l'approche théorique s'inspire de la Meaning-Text Theory [Mel'cuk 88].

La station d'Administration est conçue pour faciliter l'adaptation et la maintenance. Cette station permet à l'administrateur de modifier le style des textes à générer, en termes de structure par défaut, du titre, de l'ordre des paragraphes, des modèles de paragraphes, des types de phrases utilisées (style télégraphique / non-télégraphique), des termes utilisés, etc. En fait, il existe deux niveaux d'administration : le niveau central et le niveau local. Certaines fonctions sont spécifiques à l'administration centrale, comme la création d'un nouveau style ou la gestion de la base de terminologie multilingue. La station d'administration est écrite en Java.

## Références bibliographiques

*Coch José (1996a). « Overview of AlethGen ». Proceedings of the International Workshop on Natural Language Generation (INLG-96). Herstmonceux, Angleterre, 1996.*

*Coch José (1996b). « Evaluating and comparing three text-production techniques ». Proceedings of the 16th Conference of Computational Linguistics, Coling 96. Copenhague, Danemark, 1996.*

*Coch José, Raphaël David et Magnoler Jeannine (1995). « Quality test for a mail generation system ». Proceedings of Linguistic Engineering 95. Montpellier, France, 1995.*

*Mel'cuk Igor (1988). « Dependency Syntax : Theory and Practice ». State University of New York Press. Albany NY, USA. 1988.*

**Contact et Informations**

La page Web de MultiMeteo contient des liens avec tous les partenaires du projet ainsi qu'un prototype de démonstration :

http://www.erli.fr/erli/mm/Enmm.htm

Le projet a constitué un Groupe d'Intérêt dénommé Text Factory. Pour nous joindre, envoyez un message à :

text.factory@erli.fr

José Coch, ERLI

1, place des Marseillais

94227 Charenton-le-pont Cedex, France

Courriel : Jose.Coch@erli.fr

# Nouvelles ressources

Notes : R: pour usage de recherche - C: pour usage commercial

## ELRA-S0048 SIelex (Lexique phonétique Siemens)

Ce lexique comprend une liste de 100 000 entrées avec transcriptions phonétiques, principales marques d'accentuation et séparateurs de syllabes. La plupart de ces entrées sont extraites des rubriques politiques et économiques des quotidiens allemands *Süddeutsche Zeitung* (SZ) et *Frankfurter Allgemeine Zeitung* (FAZ). La transcription suit les grandes lignes du standard de prononciation allemande. Les exceptions sont décrites dans la documentation. Pour certaines entrées, les prononciations multiples sont prises en compte en particulier dans le cas d'homographes et d'abréviations. L'alphabet utilisé est l'extension SAM-PA à l'allemand, mais il peut être facilement traduit dans d'autres alphabets. Le jeu de caractère utilisé est ISO-8859-1. Un outil de conversion vers LATEX est également fourni avec le CD-ROM.

Prix membre : 27 500 ECU  Prix non membre : 35 000 ECU

***Offre spéciale :***

- Prix de SIelex pour les membres d'ELRA ayant déjà acquis Phonolex : 26 000 ECU
- Prix de Phonolex + SIelex pour les membres d'ELRA : 28 614 ECU

## ELRA-S0050 Base de données orale du russe (STC)

La base de données orale du russe de STC a été enregistrée entre 1996 et 1998. L'objectif principal de la base est la recherche sur la variabilité entre locuteurs et la validation d'algorithmes de reconnaissance du locuteur. Cette base a été enregistrée au moyen d'une carte son Vibra-16 16-bit de Creative Labs avec une fréquence d'échantillonnage de 11 025 Hz. Les données contiennent de la parole lue en russe de 89 locuteurs différents (54 hommes, 35 femmes), dont 70 ont enregistré 15 sessions ou plus, 10 locuteur ont enregistré 10 sessions ou plus et 9 locuteurs ont enregistré moins de 10 sessions. Les locuteurs ont été enregistrés à Saint-Petersbourg et ont entre 18 et 62 ans. Tous ont le russe pour langue maternelle. Le corpus est composé de 5 phrases. Chaque locuteur a lu attentivement mais couramment chaque phrase 15 fois à différentes dates sur une période de 1 à 3 mois. Le corpus contient un total de 6 889 occurrences et est composé de 2 volumes, pour une taille totale de 700 MB de données non compressées. Le signal de chaque occurrence est stocké sur un fichier séparé (env. 126 KB). La taille totale des données pour un locuteur est d'environ 9 500 KB. La durée moyenne d'une occurrence est d'environ 5 secondes.

Un fichier présente des informations sur les locuteurs (âge et genre). L'orthographe et la transcription phonétique du corpus est donnée dans des fichiers séparés contenant les phrases et leurs transcriptions en IPA. Les fichiers concernant les signaux sont des fichiers bruts, sans en-tête, avec 16 bit par échantillon, codage linéaire, et une fréquence d'échantillonnage de 11 025 Hz.

Les conditions d'enregistrement sont comme suit :

Microphone : microphone haute qualité, omnidirectionnel, dynamique ; distance à la bouche entre 5 et 10 cm.

Environnement : bureau.  Fréquence d'échantillonnage : 11 025 Hz.  Résolution : 16 Bit.

Carte son : Creative Labs Vibra-16.  Support électronique : CD-ROM

Prix membre : R : 400 ECU  C : 2 000 ECU  Prix non membre : R : 800 ECU  C : 4 000 ECU

## ELRA-S0051 Base de données SpeechDat(II) FDB 1000 de l'allemand

La base de données SpeechDat(II) FDB 1000 de l'allemand est composée de 988 appels à travers le réseau téléphonique fixe allemand présentés sur 4 CD-ROM. Les base orales réalisées lors du projet SpeechDat(II) ont été validées par SPEX, Pays-Bas, afin de contrôler leur adéquation avec le format SpeechDat et les spécifications de contenu.

Les éléments suivants ont été enregistrés : 1 chiffre isolé (lu ou soufflé), 1 séquence de 10 chiffres isolés, 4 chiffres connectés, 1 nombre de 4-6 chiffres permettant d'identifier la feuille de prompt, 1 numéro de téléphone d'environ 10 chiffres (lu), 1 numéro de carte de crédit de 14-16 chiffres (lu, 150 numéros différents ont été trouvés), 1 code confidentiel de 6 chiffres (lu), 1 nombre entier naturel (lu), 1 montant (argent) (lu), 3 mots épelés (1 épellation spontanée de nom, 2 lues), 1 jour (spontané), 1 phrase comportant une notion de temps (lue), 1 date (spontanée), 1 date (lue), 1 date relative (lue), 2 questions oui/non (spontanées), 3/6 mots de commande courants (lus)

Tous les mots de commande sont enregistrés plus de 80 fois. Ceux-ci sont : 1 mot de commande, 9 phrases phonétiquement riches (lues), 4 mots phonétiquement riches (lus), 5 noms provenant d'un annuaire de renseignements téléphoniques (1 nom spontané (ex : prénom), 1 nom de ville spontané, 1 nom de ville lu (provenant d'une liste de 500 noms les plus fréquents), 1 nom de compagnie lu (à partir d'une liste de 500 noms les plus fréquents), 1 nom propre, prénom et nom de famille, lu (à partir d'une liste de 150 noms).

| Prix en ECU | Membre | Non membre |
|---|---|---|
| SpeechDat(II) FDB-1000 allemand | R : 15 000/C : 18 000 | R : 25 000/C : 25 000 |
| SpeechDat(II) FDB-1000 allemand + SpeechDat(M) DB1 ou DB2 allemand | R: 20 000/C : 25 000 | R : 30 000/C : 35 000 |

***Offre spéciale*** pour l'acquisition de SpeechDat(II) FDB-1000 allemand aux membres d'ELRA ayant déjà acheté SpeechDat(M) DB1 allemand :

- Jusqu'au 30 juin 1998 : 10 000 ECU
- Entre le 30 juin 1998 et le 31 décembre 1998 : 11 000 ECU

Si l'achat de SpeechDat(II) FDB-1000 est effectué dans la même année calendaire que SpeechDat(M) DB1 ou DB2, le prix global est de :

- Prix membres d'ELRA C : 20 000 ECU/C : 25 000 ECU  Prix non membres : R : 30 000 ECU/ C : 35 000 ECU

## ELRA-L0030 Dictionnaire morphologique bulgare

Ce dictionnaire comprend 67 500 entrées réparties en 242 types de flexions (y compris les noms propres), avec une information morphosyntaxique pour chaque entrée, et un outil morphologique (MS DOS et WINDOWS 95/NT) pour l'analyse morphologique et la génération. Les données peuvent être utilisées pour l'analyse morphologique et la synthèse.

Structure des entrées : Variante linguistique locale
Format de fichier : ASCII ; lettres en minuscule
Standard utilisé : ISO
Jeu de caractères : ASCII 8-bit de code alphabétique 160-191
Support : Disquette

Prix membre d'ELRA : R : 45 ECU/C : 6 000 ECU  Prix non membre : R : 100 ECU/ C : 12 000 ECU

## ELRA-M0014 Dictionnaires bilingues (Translation Experts Ltd.)

Ces dictionnaires bilingues comportent les variantes linguistiques locales, les variantes orthographiques locales, la fréquence des mots, l'usage (familier, ancien, argot, etc.) et les traits sémantiques. Le niveau d'information varie pour chaque entrée selon le mot ou la locution et selon le dictionnaire. Cependant, toutes les informations mentionnées ci-dessus sont présentes à différents degrés dans les dictionnaires. Ces dictionnaires peuvent notamment être utilisés pour la correction orthographique, les thesauri, la troncature et la traduction en langage naturel. Un outil de traduction, également disponible à ELRA, fournit les traductions exactes, avec des sorties au format UNICODE, pour des mots et locutions entrés au format UNICODE, basées sur le vocabulaire stocké un fichier de traduction compressé.

Chaque paire de langue peut être obtenue sous la forme d'ensembles ou de sous-ensembles, correspondant au nombre d'entrées indiqué. Toutes les paires de langue sont composées de l'anglais et d'une autre langue. Les groupes de langues suivants sont actuellement disponibles :

GROUPE 1 (anglais <=> langue A) : Langue A = espagnol (25000, 60000, 100000 et 200000 entrées), français (40000, 80000, 100000 et 200000 entrées), allemand (40000, 80000 et 126000 entrées), italien (20000 et 40000 entrées), portugais-brésilien (40000, 80000 et 400000 entrées), portugais (40000, 80000, 110000 et 234000 entrées), néerlandais (40000, 80000 et 110000 entrées).

GROUPE 2 (anglais <=> langue B) : Langue B = danois (40000, 80000 et 110000 entrées), suédois (40000, 80000 et 110000 entrées), finnois (30000 entrées), islandais (40000, 80000 et 100000 entrées).

GROUPE 3 (anglais <=> langue C) : Langue C = russe (40000, 72000 et 120000 entrées), russe, domaine des affaires (60000 entrées), russe, domaine de l'aérospatiale et de l'aéronautique (60000 entrées), russe, domaine de l'automobile (40000 entrées), russe, domaine des minéraux et mines (60000 entrées), polonais (30000, 80000, 124000 et 150000 entrées), hongrois (30000, 80000 et 124000 entrées), tchèque (40000 entrées), roumain (10000 entrées).

GROUPE 4 (anglais <=> langue D) : Langue D = croate (30000 entrées), bosniaque (30000 entrées), serbe (caractères latins ou cyrilliques) (30000 entrées).

GROUPE 5 (anglais <=> langue E) : Langue E = japonais (40000 entrées).

GROUPE 6 (anglais <=> langue F) : Langue F = grec (60000 entrées).

Format de fichier : Texte
Standard utilisé : ISO
Jeu de caractères : ASCII 8-bit et UNICODE
Support électronique : CD-ROM, disquette ou téléchargement depuis le Web.
Outils connexes : Word Translator™, NeuroTran®, InterTran™, MobileTran™.

Pour plus d'informations, voir aussi http://www.tranexp.com

Le prix par entrée est présenté ci-dessous :

| | Prix membre | | Prix non membre | |
|---|---|---|---|---|
| | usage de recherche | usage commercial | usage de recherche | usage commercial |
| GROUPE 1 | 0,06 ECU | 0,25 ECU | 0,12 ECU | 0,50 ECU |
| GROUPE 2 | 0,03 ECU | 0,18 ECU | 0,06 ECU | 0,36 ECU |
| GROUPE 3 | 0,04 ECU | 0,20 ECU | 0,08 ECU | 0,40 ECU |
| GROUPE 4 | 0,04 ECU | 0,20 ECU | 0,08 ECU | 0,40 ECU |
| GROUPE 5 | 0,50 ECU | 1,00 ECU | 1,00 ECU | 2,00 ECU |
| GROUPE 6 | 0,12 ECU | 0,50 ECU | 0,24 ECU | 1,00 ECU |