Formation et recherche dans le domaine du traitement automatique des langues en contexte universitaire

Dans cette intervention, nous nous pencherons sur le rôle spécifique que jouent ou que peuvent jouer les institutions universitaires dans la mise au point de ressources liées au traitement automatique des langues (TAL).

En contexte universitaire, la recherche et le développement dans n'importe quelle discipline sont étroitement liés à l'importance accordée à la formation dans ce domaine. En effet, l'intérêt porté à un secteur se traduit par l'embauche de spécialistes (chercheurs et professeurs), la création de nouveaux cours, l'encadrement d'étudiants aux cycles supérieurs, la tenue de séminaires, etc. Ainsi, la plus grande partie de notre propos s'articulera autour des différentes activités liées à la formation en traitement automatique des langues.

Formation et universités

Aujourd'hui, la plupart des universités offrent des formations en TAL ou dans des champs connexes qui font appel à ses techniques. Nous distinguons ici deux types de formation. Les universités peuvent offrir des cours portant spécifiquement sur le traitement automatique des langues (ou la linguistique informatique), dans lesquels les étudiants se familiarisent avec les techniques symboliques et statistiques conçues pour traiter différentes composantes de la langue (phonologie, morphologie, syntaxe, sémantique et lexique). On y trouve également des cours portant sur d'autres thématiques, mais qui abordent des notions liées au TAL. Nous pensons ici à des cours sur la recherche d'information, les outils d'aide à la traduction, la terminologie ou à des cours de linguistique plus classiques, comme la lexicographie. Ces cours ne portent pas spécifiquement sur le traitement automatique des langues, mais font appel à des notions relevant de cette discipline pour expliquer des applications.

Toutefois, les modalités varient d'un établissement d'enseignement à l'autre. Nous avons choisi de décrire la situation qui règne dans un département de linguistique et de traduction, à savoir celui de l'Université de Montréal. D'une part, il s'agit de la situation que nous connaissons le mieux. D'autre part, ce département est sans conteste représentatif d'autres départements nord-américains où on enseigne la traduction ou la linguistique. Deux précisions s'imposent toutefois :

1. La formation et la recherche en TAL sont rarement confinées à un seul département universitaire. Souvent, les départements d'informatique hébergent des chercheurs intéressés directement ou indirectement par le TAL (par exemple, la traduction automatique, ou encore la recherche d'information (RI) qui intègre un nombre croissant de traitements morphosyntaxiques et sémantiques). On peut également citer d'autres secteurs qui comportent des volets qui peuvent être associés au TAL : des travaux en enseignement des langues qui visent à mettre au point des didacticiels intégrant des analyses linguistiques (par exemple, des règles de grammaire ou des représentations sémantiques) ; des travaux en sciences de l'information (par exemple, indexation automatique, génération semi-automatique de thésaurus). Par ailleurs, des collaborations peuvent s'établir entre des chercheurs affiliés à ces divers départements ;
2. La configuration du Département de linguistique et de traduction est un peu particulière en ce sens qu'elle offre deux séries de programmes bien distincts. Dans d'autres universités, la traduction et la linguistique peuvent être offertes par deux entités autonomes ou encore être rattachées à d'autres disciplines. Par exemple, la traduction est souvent offerte dans des départements de littérature comparée. La situation un peu particulière de l'Université de Montréal nous forcera à aborder les liens avec le TAL à partir des deux disciplines.

Formation TAL en linguistique

En linguistique, le traitement automatique des langues est souvent présenté comme un des domaines présentant les meilleures perspectives d'emploi. À ce titre, Montréal jouit d'une situation privilégiée, puisqu'on y trouve des entreprises travaillant dans des domaines compatibles avec le TAL : traitement automatique de la parole, traduction, correction automatique, etc. Les départements universitaires offrant des formations en linguistique ont tout intérêt à s'adapter à cette réalité. Par ailleurs, même dans des champs classiques, l'apprentissage d'applications informatiques est désormais obligatoire. Par exemple, on s'attend désormais d'un lexicographe qu'il maîtrise toute une série d'outils

logiciels, qu'il puisse manipuler des corpus de grande taille, les annoter et les interroger.

Parmi les domaines du TAL dans lesquels il sera important de former les futurs diplômés en linguistique, citons le traitement automatique de la parole (synthèse ou reconnaissance), la morphologie et la syntaxe computationnelles (les étudiants devraient pouvoir mettre au point des grammaires et des analyseurs morphologiques), la lexicographie informatisée (apprendre à enrichir ou à élaborer des dictionnaires électroniques, ou des bases de données lexicales directement utilisables dans des programmes de traitement automatique). En outre, les étudiants devraient être en mesure de faire des descriptions fines dans des langues moins souvent décrites, des langues pour lesquelles on ne trouve pas de ressources déjà construites (comme des grammaires ou des dictionnaires). Une formation en statistique linguistique semble également incontournable, même si cette dernière formation est sans doute plus approfondie dans les départements d'informatique. L'encadré 1 montre quels sont les cours offerts dans les programmes de linguistique de l'Université de Montréal. Il convient de souligner qu'actuellement, les cours de TAL sont optionnels au premier cycle.

ENCADRÉ 1
COURS DE LINGUISTIQUE AVEC UNE ORIENTATION TAL (UNIVERSITÉ DE MONTRÉAL)

- 1er cycle (cours à option): *Introduction aux langages formels, Grammaires formelles, Atelier de programmation linguistique, Traitement automatique du langage, Introduction à la lexicométrie,* quelques cours d'informatique;
- 2e cycle (cours à option): *Linguistique informatique*, *Traduction automatique, Terminologie et ordinateur* (option linguistique appliquée – Terminologie);
- 3e cycle (Ph. D. en linguistique ou option Intelligence artificielle): *Phonologie computationnelle*, *Morphologie computationnelle, Syntaxe computationnelle*, *Sémantique computationnelle.*

Quatre des 18 professeurs de linguistique du département se partagent ces cours et forment les étudiants de deuxième et de troisième cycles dans le secteur du TAL[36].

36. On trouvera d'autres détails, notamment la description des cours, dans le site Web du département: http://www.ling.umontreal.ca

Évidemment, cette initiation aux multiples facettes du TAL repose nécessairement sur une formation fondamentale en linguistique. La difficulté consiste alors à trouver un équilibre entre les domaines fondamentaux de la linguistique (phonétique, phonologie, morphologie, syntaxe, lexicologie, sémantique) et les applications spécifiques qu'on en fait en traitement automatique des langues.

En outre, les formateurs doivent s'adapter aux multiples changements que subissent les domaines du traitement automatique des langues. Par exemple, ces dernières années, les chercheurs semblent accorder une plus grande priorité à la mise au point et à la diffusion de ressources linguistiques en format électronique, c'est-à-dire des corpus textuels, des dictionnaires enrichis d'information morphosyntaxique et sémantique, au détriment d'une formation plus classique aux théories et aux formalismes linguistiques ou logiques.

Enfin, les formateurs en linguistique doivent prévoir des contenus de cours qui se démarquent des formations en TAL données dans d'autres disciplines (en informatique ou en sciences de l'information) et trouver des manières de valoriser une formation qui intègre une composante forte en linguistique.

Formation TAL en traduction

En traduction, les besoins de formation se présentent différemment. Les traducteurs sont avant tout des utilisateurs d'applications informatiques et rares sont ceux qui deviendront des concepteurs véritables. Le rôle des formateurs consiste alors à les initier à la multitude d'outils d'aide à la traduction qui orneront éventuellement leur poste de travail (bases de données, logiciels de terminologie, mémoires de traduction, aligneurs de phrases, correcteurs automatisés, fonctions avancées du traitement de texte, concordanciers, etc.).

À l'heure actuelle, toutes les universités offrant des programmes de traduction sont tenues de rendre obligatoires des cours portant sur les outils d'aide à la traduction. Les employeurs s'attendent en effet de leurs futurs employés qu'ils soient en mesure d'en faire une

utilisation efficace dès l'embauche. On offre également, depuis quelques années, des programmes de formation en localisation dans lesquels les étudiants sont appelés à se familiariser avec différentes techniques de traduction de logiciels ou de pages Web. L'Université de Montréal ne fait pas exception et offre un microprogramme en localisation qui s'adresse à des traducteurs en exercice désireux de se familiariser avec ces nouvelles techniques.

L'encadré 2 montre quels sont les cours offerts dans les programmes de traduction de l'Université de Montréal. Le cours de premier cycle intitulé *Outils informatiques des langagiers* est obligatoire ainsi que les cours de localisation (pour les étudiants inscrits au microprogramme). Les autres cours, toutefois, sont optionnels.

ENCADRÉ 2
COURS DE TRADUCTION AVEC UNE ORIENTATION TAL (UNIVERSITÉ DE MONTRÉAL)

- 1er cycle: *Outils informatiques des langagiers* (option: Initiation à la localisation et quelques cours d'informatique);
- 2e cycle: (option: *Traductique*, possibilité de suivre quelques cours de linguistique);
  DESS en traduction: *Outils informatiques des langagiers*
  Microprogramme en localisation: *Initiation à la localisation, Atelier de localisation, Projet personnel en localisation* (option: *Traductique*, cours de bibliothéconomie).
- 3e cycle: (option: *Traductique,* possibilité de suivre des cours de linguistique).

Deux des 11 professeurs de traduction du département se partagent ces cours et forment les étudiants de deuxième et de troisième cycles dans le secteur du TAL[37].

En traduction, le défi principal des formateurs consiste à intégrer dans les cours quelques notions de traitement automatique des langues et à se détacher d'un enseignement qui ressemble plutôt à une série de modes d'emploi de logiciels. Ici aussi, les formateurs doivent suivre l'évolution dans le domaine et sont sans cesse appelés à moduler le contenu des cours.

Les départements peuvent tenter d'encourager une partie des étudiants à poursuivre des études de deuxième et de troisième cycles en traductique, même si les programmes de traduction ont une forte orientation professionnelle. Les étudiants intéressés à poursuivre leurs études complètent généralement leur formation de premier cycle par des cours de linguistique formelle. Ceux qui ont fait le choix de se spécialiser en traductique peuvent suivre différents cours à option comme *Linguistique informatique* ou *Traduction automatique.* Ils ont aussi la possibilité de suivre certains cours d'informatique.

Souvent, dans les programmes de traduction, la terminologie constitue *la* porte d'entrée vers ces secteurs compatibles avec le TAL. Les cours de terminologie sont ceux où les étudiants commencent à entendre parler des dictionnaires électroniques, de corpus électroniques et à se familiariser avec les notions de base de la lexicologie.

Recherche en TAL en linguistique et traduction

En contexte universitaire, la recherche dans le domaine du TAL s'organise de différentes manières. D'abord, des laboratoires peuvent consacrer la totalité ou une partie de leurs activités au traitement automatique des langues. Par ailleurs, d'autres chercheurs peuvent mener des recherches individuelles sur un aspect ou un autre du TAL. Enfin, comme nous l'avons déjà dit plus haut, les collaborations peuvent s'instaurer entre les chercheurs d'un département de linguistique et de traduction et d'autres départements, ou encore avec des groupes extérieurs à l'université (des entreprises privées, d'autres universités canadiennes ou des universités étrangères).

À l'Université de Montréal, tous ces cas de figure sont représentés. Un premier groupe de recherche, l'Observatoire de linguistique Sens-Texte (OLST)[38], consacre une partie de ses activités à des applications relevant du TAL. Il met un

37. On trouvera d'autres détails, notamment la description des cours, dans le site Web du département: http://www.ling.umontreal.ca

38. L'adresse du site Web du groupe OLST est la suivante: http://www.fas.umontreal.ca/LING/olst/ .

accent particulier sur les modélisations lexicales et la sémantique formelle. Il consacre par ailleurs une partie de ses activités à la confection de ressources électroniques, comme des corpus généraux et spécialisés, des concordanciers pour les interroger, des corpus annotés, des ressources lexicales et terminologiques. Dans d'autres groupes de recherche, on mène des recherches dans des domaines de la linguistique compatibles avec certains domaines du TAL, comme la phonétique ou la syntaxe formelle. Enfin, des liens de collaboration sont établis entre le Laboratoire de Recherche Appliquée en Linguistique Informatique (RALI)[39] du Département d'informatique et de recherche opérationnelle sous la forme de séminaires hebdomadaires auxquels sont conviés étudiants de 2e et de 3e cycles, chercheurs universitaires et entrepreneurs[40].

Le financement de ces activités de recherche repose principalement sur des fonds octroyés par des organismes gouvernementaux, comme le Conseil de recherche en sciences humaines du Canada (CRSH), le Fonds québécois de la recherche sur la société et la culture (FQRSC), le Conseil de recherches en sciences naturelles et génie du Canada (CRSNG). Il est à noter que le gouvernement fédéral du Canada a fait du vaste domaine des industries de la langue, un domaine de développement prioritaire en 2001. Les chercheurs peuvent également obtenir des fonds auprès d'organismes internationaux comme l'Agence universitaire de la Francophonie (AUF) ou en concluant des ententes avec des entreprises privées.

La présence de groupes de recherche et l'instauration de collaborations avec des entreprises a un effet direct sur l'intérêt des étudiants pour le domaine du TAL. Les données des tableaux 1 et 2 donnent une idée de l'attrait que présente ce secteur pour la relève en linguistique et en traduction. Le tableau 1 montre que le cinquième (22 %) des travaux menés dans un département de linguistique comme le nôtre le sont dans le domaine du TAL. En traduction, la proportion est moins importante, à 10 % (tableau 2). Ce faible taux s'explique notamment au fait que la formation à l'utilisation des outils informatiques est assez récente dans notre département et dans d'autres départements québécois ou canadiens.

39. L'adresse du site Web du groupe RALI est la suivante : http://www-rali.iro.umontreal.ca/ .

40. Le programme des séminaires est affiché à l'adresse suivante : http://www-rali.iro.umontreal.ca/seminaires.html .

Tableau 1
répartition des mémoires et des thèses en linguistique (1996-2002)[41]

	Mémoires	**Thèses**
TAL	4	5
Compatibles avec TAL	8	8
Non TAL	46	42

Tableau 2
répartition des mémoires et des thèses en traduction (1996-2002)[42]

	Mémoires	**Thèses**
TAL	6	–
Compatibles avec TAL	2	1
Non TAL	53	30

Quelques pistes pour stimuler les recherches en TAL dans les universités

Nous croyons que les universités ont un rôle à jouer afin de stimuler l'intérêt des étudiants pour le domaine du traitement automatique des langues (TAL). Les programmes de linguistique et de traduction représentent souvent le premier contact des étudiants avec ce secteur.

Un premier moyen permettant de stimuler les recherches dans le domaine du TAL consisterait à donner aux étudiants de premier cycle – au Québec, les étudiants du baccalauréat – un certain nombre de cours obligatoires dans le domaine du TAL. En linguistique, la formation gagnerait à être plus *générale* : elle porte parfois sur des modèles élaborés pour régler des problèmes linguistiques hautement pointus, sans toujours prévoir un enseignement des techniques de base, comme les méthodes de traitement de chaînes de caractères ou d'élaboration de dictionnaires électroniques. Les encadrés 3 et 4 résument les quelques pistes que nous venons d'évoquer. En traduction, les formateurs pourraient intégrer davantage de notions fondamentales lorsqu'ils abordent les aides à la traduction.

41. Thèses et mémoires terminés ou en cours.

42. Thèses et mémoires terminés ou en cours.

Souvent, on a tendance à focaliser sur les applications en tant que telles, sans les placer dans un contexte plus global. L'étudiant peut devenir un expert dans la manipulation d'un produit commercial particulier, mais n'est pas toujours en mesure de généraliser des modes de fonctionnement à des familles d'applications[43].

Il conviendrait également de mieux baliser les enseignements donnés aux futurs acteurs du secteur du TAL. Souvent, la linguistique, la traduction et l'informatique semblent être des disciplines concurrentes. Il faudrait assurer, plutôt, que les formations reçues par les linguistes, les traducteurs et les informaticiens sont complémentaires.

Les suggestions faites pour bonifier la formation dans le domaine du TAL sont résumées dans les encadrés 3 et 4.

ENCADRÉ 3
FORMATION DE PREMIER CYCLE:
QUELQUES PISTES INTÉRESSANTES

- Rendre certains cours en TAL obligatoires au premier cycle;
- Enseigner certaines techniques de base en TAL aux étudiants;
- Mieux définir le rôle des linguistes et des traducteurs dans la conception d'outils de TAL;
- Souligner le rôle central des applications informatiques dans les cours plus traditionnels (ex.: traduction, terminologie, lexicographie);
- Rappeler les liens existant entre la *formation traditionnelle* et l'intégration d'applications informatiques;
- Encourager la spécialisation aux deuxième et troisième cycles seulement.

ENCADRÉ 4
FORMATION DE 2e et 3e CYCLES:
QUELQUES PISTES INTÉRESSANTES

- Former les étudiants à l'utilisation de programmes de traitement des données linguistiques (corpus, dictionnaires);
- Fournir des plates-formes de développement de ressources linguistiques formelles (analyseurs, grammaires ou entrées lexicales);
- Faire participer les étudiants à des descriptions linguistiques;
- Définir des projets de recherche débouchant sur des applications concrètes.

En recherche, quelques suggestions pourraient stimuler les activités liées au TAL. Il importerait d'abord de bien définir le rôle du chercheur universitaire en linguistique, en traduction ou en informatique. Actuellement, certains programmes de financement exigent de la part des chercheurs – pour qu'ils soient admissibles – qu'ils s'engagent à livrer un produit à la fin de leur projet. Il faut se demander si l'université est vraiment l'endroit pour faire ce genre de travail, puisqu'elle concurrence ainsi directement le secteur privé. On peut penser que le rôle du secteur universitaire consiste davantage à œuvrer en amont de l'entreprise – à contribuer à l'élaboration d'applications dans une optique à long terme, en travaillant sur des questions de recherche fondamentale.

En terminant, il importerait pour les laboratoires de faire un meilleur partage de leurs ressources. Trop souvent, les laboratoires travaillent isolément, rechignent à partager les données linguistiques dont ils disposent avec le voisin ou créent des ressources incompatibles avec les siennes.

Marie-Claude L'Homme,
Observatoire de linguistique Sens-Texte (OLST),
Université de Montréal, Québec.

43. Quelques solutions à ce problème sont proposées dans le cadre du cours Outils informatiques des langagiers. Voir Robichaud, Benoît et Marie-Claude L'Homme (2003), «Teaching the Automation of the Translation Process to Future Translators», dans Workshop on Teaching Translation, Technologies and Tools, XV Machine Translation Summit, 23-27 septembre 2003 et L'Homme, Marie-Claude (2003), «Traduction et outils informatiques: mise en forme d'un cours d'initiation», dans Mareschal, Geneviève et autres (éd.), La formation à la traduction professionnelle, Ottawa: Les Presses de l'Université d'Ottawa, pp. 177-197.