TAL et Langues peu dotées

Chantal Enguehard
Laboratoire d'Informatique de Nantes Atlantique – Université de Nantes
2, rue de la Houssinière
BP 92208 44322 Nantes Cedex 03 France
chantal.enguehard@univ-nantes.

Ces dix dernière années ont vu de grands bouleversements dans l'accès aux Nouvelles Technologies. La Toile s'est considérablement étendue et recouvre maintenant la planète,touchant la quasi-totalité des cultures. Cette extension spatiale a été accompagnée par une capacité accrue à représenter les différentes langues. Alors que les caractères étaient codés grâce à un seul octet dans la représentation ASCII, le standard Unicode, apparu en 1992,définit une représentation sur 4 octets qui permet de représenter de manière unique chacun des caractères de chacune des langues. Désormais, le stockage des documents sous une forme électronique qui permet leur traitement analytique autorise de nombreuses langues à franchir la première étape de l'informatisation.

Ce progrès considérable doit être soutenu. Même si le standard Unicode tend à représenter les caractères de toutes les langues, l'inventaire de ces caractères n'est pas entièrement complété,ou bien ce standard de représentation est encore peu connu, et donc peu respecté. Deux articles détaillent ces difficultés :

- Grégory Kourilsky présente la cas de l'écriture tham du Laos qui n'a pas de représentation dans Unicode. Il émet des propositions détaillées pour remédier à cette situation.

- Wunna Ko Ko et Mikami Yoshiki inventorient huit langues du Myanmar, certaines écrites depuis des siècles. Bien que ces langues utilisent majoritairement des caractères représentés dans Unicode, leur informatisation souffre du manque de standardisation (le développement de polices de caractères non normalisées gène le partage des textes) et de coopération entre les chercheurs.

Les travaux visant à constituer des ressources linguistiques sont souvent insuffisants : soit il y a eu peu de recherches en linguistique, soit celles-ci n'ont pas produit de ressources électroniques utilisables. La constitution de telles ressources est une première étape cruciale pour fonder des travaux en Traitement Automatique des langues. Il s'agit de constituer des corpus de textes de taille importante afin d'en extraire des ressources linguistiques électroniques.

- Hubert Naets s'appuie sur un échantillon de langue pour constituer automatiquement un corpus à partir de la Toile. Cette procédure statistique permet de distinguer finement les langues voisines.

- Daniel Yacob présente les travaux menés pour constituer un corpus de l'Amharic et en déduire un lexique avec une représentation normalisée. Il détaille les aspects légaux rencontrés lors de la collecte de textes.

- Emmanuel Schang expose les apports d'un corpus oral transcrit pour saisir une langue dans son expression spontanée.

- Claudia Soria et Monica Monachini s'appuient sur des corpus en différentes langues (dont une langue linguistiquement bien dotée), et traitant d'un même domaine, pour extraire automatiquement la terminologie de ce domaine dans chacune des langues.

L'adaptation de travaux existants, la mise au point de stratégies facilement adaptables à différentes langues constituent en enjeu important pour équiper les langues en outils automatiques.

- Laurent Besacier, Viet-Bac Le, Eric Castelli, Sethserey Sam et Ludovic Protin cherchent à adapter un système de reconnaissance automatique de la parole continue à deux langues peu dotées : le vietnamien et le khmère.

- Johannes Heinecke vérifie qu'il est possible d'adapter un étiqueteur morphosyntaxique au gallois même si cette langue présente des caractéristiques supplémentaires par rapport à la langue initiale visée par l'étiqueteur.

- Frédérick Houben et François Rioult s'appuient sur des propriétés très générales des langues et des méthodes de fouilles de données pour effectuer automatiquement l'étiquetage de textes.

Il apparaît cependant que l'étude linguistique fine d'une langue est indispensable lors de certaines étapes.

- Les travaux de Bali Ranaivo-Malançon visant à construire un étiqueteur morphosyntaxique du malais en s'appuyant sur la morphologie des mots de cette langue se heurtent au manque de consensus sur le jeu d'étiquettes adéquats à utiliser pour cette langue.

Enfin, nous avons inclus un article ne traitant pas directement du Traitement Automatique des Langues mais de la localisation des logiciels. En effet, si les utilisateurs souhaitent s'exprimer dans leurs langues et bénéficier de l'apport d'outils automatiques, comme les correcteurs orthographiques par exemple, ils apprécient également que les outils électroniques s'adressent à eux en respectant leurs langue et culture.

- Dawit Bekele explique l'importance de la localisation des logiciels pour les pays du tiers-monde, puis il détaille les différents aspects techniques de la localisation.

Les recherches présentées lors de cet atelier abordent différents points de la chaîne des traitements appliqués aux langues. Elles constituent également un apport plus général pour le Traitement Automatique des Langues puisqu'elles soulignent des difficultés que peuvent rencontrer toutes les langues, y compris les langues largement dotées comme l'anglais, le français ou l'espagnol. En effet, les langues évoluent, des expressions singulières apparaissent Atelier TAL et Langues peu dotées et il faut trouver des stratégies pratiques pour faire évoluer conjointement les ressources linguistiques sur lesquelles s'appuient les traitements automatiques.