Outils de traitement de corpus et langues peu dotées

Marguerite Leenhardt - M2P Ingénierie Multilingue INaLCO

12 janvier 2008

Résumé

Nous nous proposons ici de donner un panorama des recherches et outils développés pour le traitement automatique des langues peu dotées, ou langues-π. La possibilité de pouvoir constituer et analyser des ressources linguistiques informatisées pour de telles langues relève de perspectives humaines et culturelles. En effet, les travaux de recherche du TAL appliquées aux langues minoritaires contribuent sans doute de leur survie, s'inscrivant dans une perspective de préservation du patrimoine culturel associé à une langue, et peuvent par ailleurs constituer paramètre décisif dans l'alphabétisation des populations.

Introduction

(Berment, 2004) définit dans sa thèse la notion de *langue peu dotée*, dont le niveau d'informatisation peut être défini à l'aide d'un indice σ, calculé comme suit : *A chaque service ou ressource, un groupe d'utilisateurs représentatifs des locuteurs de la langue attribue un niveau de criticité C_k et une note N_k, la moyenne pondérée des notes - appelée* ***indice-σ*** *- reflétant leur satisfaction globale.* Une langue-π est déterminée par une moyenne pondérée inférieure à 10. Après avoir abordé le problème du codage des caractères, préalable nécessaire à toute constitution de ressources linguistiques et donc d'outils de traitement de corpus, nous présentons des travaux relatifs à la constitution de corpus pour les langues peu dotées. Nous parlons ensuite du transfert de techniques existantes pour des langues bien dotées, en particulier concernant les outils de traitement de l'oral et de l'écrit. Nous évoquons également le problème de la localisation des logiciels, qui, bien qu'il ne s'inscrive pas, à proprement parler, dans une perspective de TAL, a néanmoins sa place du point de vue de l'adaptation de techniques aux langues-π.

1 Codage des caractères

Le codage des caractères est le premier pas vers la constitution de ressources électroniques exploitables automatiquement. En amont du problème du développement d'outils pour traiter une langue donnée, il faut se poser celui de la disponibilité de données numérisées dans cette langue. Cela suggère de définir pour les langues dont les systèmes d'écritures ne sont pas similaires aux langues numérisées, un codage adapté. L'amharique, objet des travaux de (Yacob, 2005), en est un excellent exemple : *given the encoding difficulties there was little expectation that a document written on one computer should be readable on another.*
Les travaux de (Kourilsky,2005) présentent une remise en question des modèles de stockage recommandés par le consortium Unicode, qui, s'ils sont optimaux pour des tâches de traitement automatique, sont difficiles d'accès pour les utilisateurs. En effet, les méthodes de rendu intégrées par Unicode, bien qu'elles permettent

une résolution des difficultés inhérentes au traitement automatique des écritures appartenant à la famille des écritures indiennes, qui sont non linéaires et non connexes, impliquent cependant une mise en oeuvre complexe qui en pénalise l'utilisation. Il présente une méthode de codage proche de l'écriture manuscrite, dans laquelle les signes non linéaires et non connexes associés aux consonnes pour former un mot sont considérés comme des caractères, s'inspirant des protocoles de codage du modèle thaï-lao, antécédent à Unicode. Au-delà des considérations techniques, ce type de travaux s'insère dans une problématique plus générale qui tend à considérer qu'élever le niveau d'informatisation d'une langue peu contribuer à lutter contre l'analphabétisme, qui touche majoritairement les pays dont les langues sont faiblement dotées informatiquement.

2 Constitution de corpus

Comme le souligne (Besacier, 2006), *de nombreuses langues du monde sont essentiellement orales et n'ont pas de forme écrite répandue.* Il est couramment admis que seules 10% des langues parlées dans le monde utilisent l'un des 25 systèmes d'écritures connus à ce jour. Dans ce cas, l'accès aux données pour le traitement automatique est rendu difficile, puisque le seul moyen d'accéder à des ressources textuelles est de constituer des corpus d'oral transcrit. La collecte de corpus est pourtant essentielle à la constitution de ressources linguistiques et au développement d'outils de traitement de corpus.

L'utilisation des données multimédia diffusées sur l'Internet est une des stratégies les plus répandues dans le cadre de la constitution de ressources pour les langues-π. En effet, pour ne citer qu'eux, les médias d'information nationaux ou régionaux ont le plus souvent une version papier numérisée [1] et les données de type radio-télévision diffusées (broadcast news) sont nombreuses. Cependant, certaines langues, comme par exemple l'amharique, ont une visibilité réduite sur le web. Le problème de collecte des données monolignues peut également se poser pour les pays anciennement colonisés, comme l'illustre le cas du malgache, dont le contenu des pages web est majoritairement bilingue (français-malgache), voire intégralement en français.

(Scannel, 2007) estime qu'environ une centaine de langues, hormis les quelques dizaines de langues dont le niveau d'informatisation est élevé, sont parvenues à se doter de ressources linguistiques élémentaires, telles les corpus monolingues et bilingues, dictionnaires électroniques, thésaurus, analyseurs morpho-syntaxiques et autres étiqueteurs. Ses travaux présentent un *crawler* web dont l'objectif est la récolte automatique de données pour les langues peu dotées, développé dans cadre du Crudaban Project[2]. Cette initiative a permis la mise à disposition de ressources variées, par exemple des corpus de phrases et des ressources pour l'identification automatique des langues rares, distribuées et exploitées dans le cadre d'autres projets allant de la lexicographie à la traduction automatique, en passant par la désambiguïsation et la constitution de thésaurus. (Scannel, 2007) prolonge ainsi des travaux du même type initiés par (Ghani et al., 2003) et (Naets, 2005), qui ont eux aussi développé des crawlers web ayant pour objectif la constitution de corpus de langues minoritaires. Ces systèmes ont en commun non seulement l'application d'une procédure statistique et l'utilisation d'un échantillon de langue pour leur tâche de constitution automatique de corpus, mais aussi une exploitation maximale de la proximité entre les langues pour générer des requêtes

[1] Cf. par exemple http ://www.mediatico.com/fr/
[2] http ://borel.slu.edu/crubadan/

envoyées à un moteur de recherche, ainsi que l'intégration d'un module d'identification automatique des langues des textes retournés par une requête donnée.
S'il est évident que les faits historico-géopolitiques ont une incidence majeure sur l'utilisation, l'enseignement et la diffusion des langues minoritaires, les politiques linguistiques nationales doivent également être prises en compte lorsque l'on s'atèle à une tâche de constitution de corpus pour une telle langue. (Yacob, 2005) présente un panorama des travaux de recherche ayant pour objectif commun la constitution de corpus en amharique, et les initiatives mises en oeuvre pour la constitution d'un lexique normalisé. La complexité de l'orthographe amharique, et les difficultés de codage qui en découlent, représente en soi une entrave à l'informatisation des données, en particulier au développement de corpus, de lexiques et d'un standard de transcription. En marge de ces considérations, il fait une place importante dans son exposé aux aspects législatifs, dont les changements en Ethiopie ont pu et peuvent représenter une gêne, sinon un obstacle, aux initiatives de développement de ressources accessibles publiquement.
Les travaux de (Nimaan et al.,2006) répondent quant à eux aux problèmes d'accès aux données et de sauvegarde du patrimoine culturel pour les langues de pays à tradition orale, en particulier la langue somali parlée à Djibouti et en Afrique de l'Est. Différents aspects sont pris en compte parmi les outils développés, dans un but d'indexation des données orales et de leur transcription automatique, en particulier, l'adaptation de techniques de constitution de corpus, à partir du web, par synthèse automatique de texte et par traduction automatique, pour pallier le manque de corpus existants.
L'un des aspects abordés par (Soria et Monachini, 2005) est relatif à l'élaboration d'une méthodologie de production de ressources linguistiques, en particulier pour les balkaniques et les langues de l'Europe de l'Est, telles le russe, le polonais, l'ukrainien, le roumain, le serbo-croate, mais aussi le grec, le serbe, le bulgare et le slovène. Ces travaux mettent en exergue la nécessité de constituer des ressources dans ces langues qui soient facilement accessibles, afin que la couverture des langues occidentales en termes de données numériques disponibles soit plus équilibrée, et ne reflète pas l'état de fait actuel d'une europe à deux vitesses, circonstance dans laquelle les langues de l'ouest sont largement avantagées.

3 Transfert de techniques

3.1 Des stratégies adaptables

Les outils de traitement de l'oral comptent parmi l'ensemble des critères spécifiés par (Berment, 2004), permettant d'évaluer qu'une langue est peu dotée informatiquement. C'est pourquoi (Besacier et al.,2005) présentent dans leurs travaux une méthodologie visant à développer et adapter un système de reconnaissance automatique de la parole continue, en particulier pour le vietnamien. En effet, si de nombreux outils ont été mis au point pour le traitement textuel de cette langue, dont la transcription en alphabet latin enrichi d'accents est peu problématique, les outils de traitement de l'oral sont quasi inexistants. Le recueil de données textuelles, qui sont nécessaires au développement d'un système de reconnaissance automatique de la parole, ne constitue pas un obstacle majeur. Cependant, la constitution de ressources pour la définition d'un modèle acoustique est généralement coûteux. Le choix de (Besacier et al.,2005) pour la constitution d'un dictionnaire phonétique mobilise une analyse phonétique par règles d'un lexique de mots isolés issus d'un dictionnaire bilingue français-vietnamien. Les performances du système sont pénalisées par la mise en correspondance de systèmes phonologiques qui, bien que

présentant de nombreux sons communs, sont cependant différents.
(Pellegrini,2005) situe ses travaux dans une perspective analogue, puisqu'il se place dans le cadre de la transcription automatique de langues rares, en particulier la langue amharique. Il s'agit d'évaluer l'influence des méthodes de décomposition sur les performances d'un tel système. La décomposition des mots a en effet une incidence fondamentale sur la couverture des lexiques, la taille des textes d'apprentissage et l'élaboration de modèles de langage, puisque la transcription automatique à partir de morphèmes obtient de meilleures performances qu'avec des mots, du point de vue de la mesure de perplexité, utilisée pour évaluer les modèles de langage. Apporter des éléments de solutions à des problématiques très spécifiques, comme c'est ici le cas, participe de l'adaptation de techniques existantes visant à contribuer à la dotation en outils de traitement de corpus pour les langues-π.
(Somers et al., 2005) présentent un système de synthèse vocale ou *text-to-speech* (TTS) utilisant un système existant pour une langue dotée (langue source) qui simule du TTS pour une langue-π (langue cible), dans le but de fournir un soutien à des locuteurs pour lesquels la barrière linguistique peut représenter un désavantage majeur, par exemple des immigrants ne pratiquant pas la langue de leur terre d'accueil. Il s'agit d'identifier des facteurs permettant de choisir une langue source adéquate pour générer du TTS en langue cible, en particulier pour la langue somali. Une telle démarche nécessite d'adapter un outil existant à une langue peu dotée, qui présente un système phonologique et des caractéristiques prosodiques analogues à la langue cible à générer, de telle sorte que pour une situation d'interaction mettant en présence deux locuteurs de langues différentes, le système puisse générer une traduction permettant au locuteur de la langue cible une compréhension générale de l'énoncé qui lui est adressé. (Somers et al., 2005) adaptent un synthétiseur vocal allemand pour générer du somali ; les problèmes majeurs pour la compréhension sont phonétiques et lexicaux, les tons étant sémantiquement pertinents en somali. Cependant, la synthèse vocale reste compréhensible. Ces travaux se situent donc eux-aussi dans la perspective de transfert de techniques pour le développement d'outils adaptés aux langues-π.

Les outils de traitement de l'écrit comptent également parmi les critères de l'évaluation du niveau d'informatisation d'une langue donnée.
L'amazighe (berbère) est une langue peu dotée informatiquement, récemment intégrée par le standard Unicode[3], ce qui a permis le développement d'outils adaptés au traitement de cette langue. En particulier, (Rachidi et Mammass, 2005) discutent des méthodes pour la mise en oeuvre d'un traitement de texte et le développement d'un système de traduction automatique et de gestion d'un base lexicale amazighe, et présentent la réalisation d'un traitement de texte amazighe.
Dans la perspective de sauvegarde du patrimoine culturel pour les langues de pays à tradition orale, en particulier la langue somali parlée à Djibouti et en Afrique de l'Est, les travaux de (Nimaan et al.,2006) relèvent de l'adaptation de techniques pour le développement d'outils pour la racinisation, la conjugaison, la transduction, la phonétisation et le résumé automatique, visant à rendre exploitable l'état écrit des corpus oraux.
(Houben et Rioult, 2005) et (Houben et Rioult,2006) présentent quant à eux les prémisses du développement d'une méthode procédant par apprentissage supervisé, dont l'objectif est d'obtenir une catégorisation des mots à partir de propriétés accessibles depuis le corpus brut. Il s'agit d'attribuer une étiquette

[3] L'amazighe a été intégrée à Unicode en 2004.

morphosyntaxique à un mot en fonction de sa position relative aux autres mots du texte. Cette méthode, contrairement à la majorité des étiqueteurs, n'utilise pas de lexique, pas de corpus étiqueté et ne fait pas appel aux règles symboliques pour déterminer les catégories. A partir d'un ensemble d'attributs des mots (mots de type pleins/vides, succession de ces types dans le texte, position du mot dans le virgulot, influence d'un mot vide sur la flexion des mots du cotexte), une classification supervisée est appliquée, corrélée à une recherche de conjonction d'attributs des mots. Ces travaux fonderaient une méthode de validation des propriétés des mots présentant l'avantage de ne pas nécessiter de ressources, qui serait donc, à ce titre, doublement avantageuse pour les langues peu dotées, puisqu'indépendante du développement préalable de ressources linguistiques, cette méthode serait non seulement très adaptable, mais en plus peu coûteuse.
(Besacier, 2006) fait état d'un analyseur morphologique développé dans le contexte de ses recherches en reconnaissance automatique de l'arabe dialectal irakien. Compte-tenu de la différence morphologique entre le dialecte irakien et l'arabe standard, le développement d'un tel outil s'est avéré nécessaire pour segmenter les données d'apprentissage du modèle de langage irakien. Ce type de modélisation morphologique a un grand intérêt à plusieurs niveaux. En premier lieu, il permet une amélioration des performances par rapport à un modèle de mots classiques. Il permet également une réduction de la taille du vocabulaire, ce qui est particulièrement pertinent pour l'implémentation d'algorithmes sur des terminaux légers ayant de faibles ressources mémoire. Enfin, cette modélisation morphologique a été implémentée avec succès dans un prototype de traduction de parole, développé dans le cadre du projet TRANSTAC[4]. Du point de vue de l'adaptation de techniques pour le traitement des langues peu ou mal dotées, dont la visibilité est réduite en termes d'accessibilité des données numériques, un autre des aspects abordés par les travaux de (Soria et Monachini, 2005) relève de la constitution de corpus parallèles multilingues et de lexiques terminologiques pour les langues balkaniques et d'Europe de l'Est. Leurs travaux s'inscrivent dans le cadre de l'INTERA Project[5], qui vise la production d'un modèle pour la construction de lexiques terminologiques multilingues. La stratégie retenue utilise une seule langue-pivot pour l'extraction terminologique dans les données des langues-cibles, en procédant par adaptation d'une méthode hybride combinant techniques statistiques et symboliques.

3.2 La localisation des logiciels

Bien qu'ils ne s'inscrivent pas à proprement parler dans la perspective du TAL, les travaux concernant la localisation des logiciels méritent ici leur place, dans la mesure où ils participent de la facilitation de l'accès par les utilisateurs aux outils de traitement automatique développés pour leur langue. Si les utilisateurs souhaitent bénéficier d'outils, tels les correcteurs orthographiques par exemple, il est fort appréciable de leur point de vue que ceux-ci soient adaptés à leurs culture et langue. (Bekele, 2005) insiste en effet sur trois motivations fondamentales relatives à l'importance de la localisation pour les pays du tiers-monde. En premier lieu, offrir à ces pays la possibilité de travailler dans leur langue officielle. En second lieu, permettre à la population jeune d'avoir accès aux NTIC, considération qui se situe dans le perspective de la dotation de ces langues comme support à la lutte contre l'analphabétisme et comme outil d'instruction. En-

[4] http ://transtac.mitre.org/

[5] http ://www.mpi.nl/INTERA/project/project.html

fin, limiter et pallier la visibilité réduite de ces langues en termes d'accessibilité aux données numérisées.

Conclusion

Certaines langues, comme le malais, n'entrent à priori pas dans le paradigme des langues peu dotées. En effet, les données numérisées disponibles sur le web pour cette langue sont nombreuses, à la différence du somali par exemple. Cependant, l'article de (Ranaivo-Malançon, 2005) soulève un problème important quant aux outils de traitement automatique dédiés à cette langue, qui permet de compter le malais parmi les langues faiblement dotées. Bien que des travaux aient été menés pour le développement d'outils de traduction automatique, de traitement de l'oral ou encore de dictionnaires multilingues intégrant le malais, *les outils pour le traitement automatique du malais sont non réutilisables car implémentés sur des systèmes désuets et utilisables uniquement pour une seule application.* Il apparaît donc impératif que les iniatives futures pour le développement d'outils destinés au traitement des langues peu dotées évitent l'écueil d'une mauvaise gestion des ressources et des outils. Par ailleurs, cet article fait écho aux problèmes posés par les politiques nationales liées à la gestion de la langue : sans normes grammaticales et orthographiques établies par les institutions qui en ont la charge, les travaux de recherche ne peuvent avoir de référence.
Si le gallois n'entre pas non plus dans un tel paradigme, - les travaux de (Heinecke, 2005) dressant un panorama des ressources linguistiques électroniques disponibles pour cette langue-, cet article souligne cependant que *le nombre de ressources reste limité.* De fait, l'auteur insiste sur la nécessité de poursuivre l'effort de développement d'outils de traitement destinés à la langue galloise. Une telle considération doit être étende aux langues pour lesquelles la visibilité en termes d'accessibilité aux données numériques n'est plus une entrave.

Enfin, si l'on admet que la notion de langue-π intègre tout système linguistique ayant peu ou pas fait l'objet d'une informatisation, la langue des signes doit être prise en compte dans ce paradigme. De fait, des initiatives telles que la mise en place de l'atelier TALS 2005, dont (Braffort et al., 2005) décrivent les thèmes et travaux de recherche sur la langue signée, sont l'indice d'une volonté de la communauté du TAL et, plus généralement, de la recherche en Linguistique Appliquée, de s'affranchir des sentiers battus des langues monopolisant la toile et dont les données numériques sont déjà disponibles en quantité.

Références

(Bekele, 2005) D. Bekele, *Localization in the context of a third world country*, Departement of Computer Science Addis Ababa Uiversity Ethiopia, TALN 2005

(Berment, 2004) V. Berment, *Méthodes pour informatiser des langues et des groupes de langues 'peu dotées'*, Grenoble, Thèse de doctorat Université Joseph Fourier Grenoble 1, 2004

(Besacier, 2006) L. Besacier, *Transcription enrichie de documents dans un monde multilingue et multimodal*, Habilitation à diriger les recherches, Université Joseph Fourier Grenoble 1, 2006

(Besacier et al.,2005) L. Besacier et al., *Reconnaissance automatique de la parole pour les langues peu dotées : application au vietnamien et au khmer*, TALN 2005

(Braffort et al., 2005) A. Braffort et al., *Atelier Traitement Automatique des*

Langues des Signes TALS 2005, TALN 2005

(Ghani et al., 2003) R. Ghani et al., *Building minority language corpora by learning to generate web search queries*, Knowledge and Information Systems, 2003

(Heinecke, 2005) J. Heinecke, *Aspects du traitement automatique du gallois*, France Télécom, Division Recherche & Développement, TALN 2005

(Houben et Rioult, 2005) R. Rouben et F. Rioult, *Généralisation d'étiquetage morpho-syntaxique par classification supervisée*, TALN 2005

(Houben et Rioult,2006) R. Rouben et F. Rioult, *Etiquetage morpho-syntaxique par classification supervisée : vers une alternative aux dictionnaires ?*, JADT 2006

(Kourilsky,2005) G. Kourilsky, *Exemple d'écriture ignorée par Unicode : l'écriture tham du laos*, INaLCO, TALN 2005

(Naets, 2005) H. Naets, *La Déclaration Universelle des Droits de l'Homme : 329 langues pour la constitution automatique de corpus et de lexiques*, Laboratoire d'Ingénierie de la Connaissance Multimédia Multilingue (LIC2M), TALN 2005

(Nimaan et al.,2006) A. Nimaan et al., *Boîte à outils TAL pour des langues peu informatisées*, JADT 2006

(Pellegrini,2005) T. Pellegrini, *Considérations pour la transcription de langues rares*, LIMSI-CNRS, TALN 2005

(Rachidi et Mammass, 2005) A. Rachidi et D. Mammass, *Vers un système d'écriture informatique amazighe : méthodes et développements*, RECITAL 2005

(Ranaivo-Malançon, 2005) B. Ranaivo-Malançon, *Approche pour un étiquetage morphosyntaxique du malais*, Unit Terjemahan Melalui Computer Universiti Sains Malaysia, TALN 2005

(Scannel, 2007) K.P. Scannel, *The Crudaban Project : corpus building for under-resourced languages*, Department of Mathematics and Computer Science, Saint Louis University Missouri, in *Cahiers du Cental, 5 (2007), 1*

(Somers et al., 2005) H. Somers, G. Evans et Z. Mohamed, *Developing speech synthesis for under-resourced languages by 'faking it' : an experiment with somali*, School of Informatics University of Manchester, TALN 2005

(Soria et Monachini, 2005) C. Soria et M. Monachini, *Methods, models and Standardization Issues for the creation of linguistic resources : the case of under-represented languages*, TALN 2005

(Yacob, 2005) D. Yacob, *Developments towards an electronic ahmaric corpus*, Ge'ez Frontier Foundation, TALN 2005