Linguistique informatique et linguistique de terrain

Claire Gardent s'intéresse aux modèles hybrides statistiques/symboliques pour la modélisation informatique des langues naturelles. Elle travaille sur l'acquisition automatique de grammaires et de ressources lexicales, sur l'analyse syntaxique et sémantique, sur la génération automatique de textes et sur les technologies innovantes pour l'apprentissage des langues. Denis Paperno étudie la sémantique de la langue naturelle, en particulier sur le problème de la sémantique des mots ou des phrases dans le traitement automatique des langues. Tous deux sont chercheurs CNRS au Laboratoire lorrain de Recherche en Informatique et ses Applications (LORIA, UMR 7503, CNRS / Université de Lorraine / Inria).

Steven Bird et Augustine : enregistrement d'une histoire en Tembé avec l'application Aikuma

7099, c'est selon Ethnologue le nombre de langues actuellement parlées dans le monde. Mais, comme le panda ou le saumon sauvage d'Écosse, beaucoup de ces langues sont en voie d'extinction. Un quart environ de ces langues sont parlées par moins d'un millier de personnes et on estime à environ 3000 le nombre de langues qui vont disparaître au cours du siècle à venir.

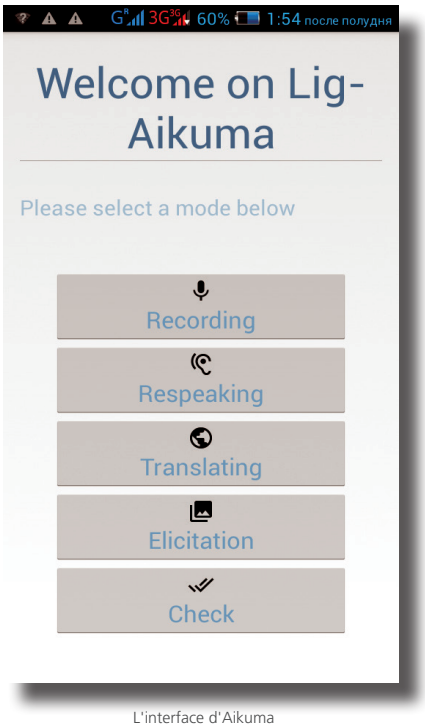
Si, pour l'anthropologue, chaque langue reflète la culture d'un groupe et contient des indications qui, combinées avec les données génétiques, permettent de tracer l'histoire des populations, pour le linguiste, chaque langue est une source précieuse d'information dans la quête qui vise à identifier l'étendue des possibles linguistiques. Inversement, chaque langue qui disparaît est un élément de moins dans l'étude qui vise à identifier les mécanismes langagiers. Il est ainsi crucial de faciliter la collecte et la documentation de ces langues en péril ainsi que des langues minoritaires qui ne sont pas en danger imminent mais subissent des changements inévitables dans le contexte de la globalisation.

Ces dernières années, la recherche en linguistique informatique apporte des éléments intéressants pour répondre à ces besoins. Dans ce qui suit, nous illustrons à l'aide d'exemples comment la linguistique informatique peut contribuer à documenter les langues en péril et plus généralement, comment elle peut contribuer à faciliter le travail des linguistes et des linguistes de terrain.

Faciliter la collecte des données

Afin de documenter les langues en voie de disparition, un besoin particulièrement urgent concerne la collecte et la traduction de données orales. Pour faciliter, accélérer et améliorer ce processus, différents outils informatiques ont récemment été proposés par des linguistes informaticiens. Ainsi, l'application mobile AIKUMA[1] développée par Steven Bird permet d'enregistrer la parole spontanée ainsi que la traduction et la « re-dite », à un rythme plus lent, de ces enregistrements[2]. Elle permet également de créer un alignement entre un enregistrement et sa re-dite et/ou sa tra-

1. Bird S., Gawne L., Gelbart K. and McAlister I. 2014, *Collecting Bilingual Audio in Remote Indigenous Communities*, COLING, https://www.aclweb.org/portal/
2. La « re-dite » vise à faciliter la transcription *a posteriori* des données enregistrées.



L'interface d'Aikuma

duction. Les données collectées sur les Smartphones étant partagées sur un réseau wifi local, tout utilisateur connecté peut en outre participer au travail de collecte en fournissant une traduction, un enregistrement ou une redite. Enfin, l'extension LIG-AIKUMA développée par le Laboratoire d'informatique de Grenoble (LIG) inclut un mode « Correction » qui permet au linguiste de corriger du texte (erreurs orthographiques, syntaxiques, de prononciation, etc.) et un mode « Collecte (élicitation) » permettant d'éliciter de la parole auprès du locuteur au moyen de textes, d'images ou encore de vidéos.

Exploitant la puissance et la légèreté des téléphones portables, ces logiciels libres de droit permettent, d'une part, de collecter des données orales de bonne qualité et, d'autre part, d'associer ces données à des textes numériques (traduction, transcription) directement utilisables par des processus informatiques en aval comme, par exemple, l'alignement texte/parole, mais également, dans le cas où la taille des données est suffisante, la détection automatique ou semi-automatique des éléments constitutifs d'une langue (phonèmes, morphèmes, mots, grammaire, etc.) Plus généralement, ils permettent non seulement d'accélérer la collecte des données pour les langues en voie de disparition mais également de produire des données de meilleure qualité pour l'analyse linguistique.

Induction de la phonologie d'une langue par des méthodes statistiques

Dans le traitement automatique des langues, les méthodes statistiques ont de plus en plus de succès et d'importance. Généralement, ces méthodes se basent sur l'utilisation de grandes quantités de données enrichies avec des annotations manuelles (par exemple, la bande audio d'un texte parlé sera annotée avec le texte écrit correspondant). Si, pour les langues comme l'anglais et le français, les données sont en effet abondantes (dizaines ou centaines d'heures de parole), ce n'est pas le cas des langues en voie d'extinction. Certaines méthodes statistiques peuvent néanmoins être appliquées avec succès. Ainsi, les participants des "*Zero Resource Speech Challenge* 2015 et 2017" (organisé par des chercheurs français, espagnols et américains) présentent des systèmes d'identification d'unités lexicales dans le langage parlé et d'identification des traits distinctifs des sons[3] qui, contrairement aux systèmes précédents, sont appris à partir d'une petite quantité de données (moins de cinq heures de parole enregistrée) et sans aucune annotation. Même s'ils ont un taux d'erreur d'environ 10 %[4], ces systèmes peuvent fortement faciliter aussi bien l'élaboration des données audio des langues sans tradition écrite que l'analyse de leurs propriétés phonétiques ce qui, typiquement, correspond à plusieurs semaines de travail manuel pour le linguiste de terrain. En effet, l'analyse d'une minute de parole enregistrée demande environ une heure de travail pour un linguiste qualifié travaillant avec un locuteur de la langue. En permettant de réduire substantiellement le temps requis par la pré-segmentation du flux oral en unités correspondant à des mots, ces systèmes automatiques facilitent la préparation de données qui peuvent ensuite être exploitées pour entraîner un système automatisé de transcription et ainsi accélérer la phase de transcription (il suffit de vérifier et, au besoin, corriger les transcriptions automatiques plutôt que de les rédiger entièrement manuellement).

Exploitation de méthodes statistiques pour l'induction d'informations lexicales

Ces dernières années, l'analyse distributionnelle est devenue un outil standard de la linguistique informatique. Cette méthode permet de représenter le sens d'un mot par une séquence de chiffres (formellement, un vecteur) indiquant les mots qui lui sont fréquemment associés dans un corpus donné. Ces vecteurs permettent notamment de déterminer des relations lexicales. Par exemple, la distance entre les vecteurs de deux mots (disons *pomme* et *arbre*) peut être exploitée pour déterminer leur degré d'association sémantique. Même si, d'habitude, on utilise de très grands corpus de textes (plusieurs centaines de millions de mots) pour obtenir des vecteurs de bonne qualité, les données limitées des langues minoritaires peuvent elles aussi être utilisées pour certaines tâches comme l'induction automatique de classes lexicales qui peuvent être soit les parties de discours traditionnelles soit des classes plus fines telles la classe des verbes de mouvement ou des prépositions locatives.

Les vecteurs de mots créés pour les langues majeures peuvent également être exploités pour analyser la variation lexicale à travers les langues du monde. Par exemple, une barbe, une conversation et une sauce peuvent toutes être *piquantes*, mais l'adjectif *piquant* se traduira différemment pour chacun de ces cas (en anglais, *prickly beard*, *racy conversation*, et *spicy sauce*). Le contraste fort entre les différents usages du mot piquant est évident si on construit les vecteurs des phrases correspondantes, sans qu'il soit besoin ni de les traduire en anglais ni d'analyser leur sémantique manuellement[5]. À partir de ces vecteurs, il est ainsi possible de construire une liste de phrases sémantiquement diverses qui pourra être utilisée pour la création d'un questionnaire lexical visant l'étude du lexique d'une langue minoritaire. Notons que ce type de questionnaire peut être créé automatiquement et de manière systématique.

3. Les traits distinctifs sont les attributs des sons qui peuvent distinguer les mots, par exemple le trait de sonorité qui oppose parmi les autres les sons [p] et [b] et distingue les mots *pas* et *bas*, *pois* et *bois*, etc.
4. Versteeg M., Thiollière R., Schatz T., Cao X., Anguera X., Jansen A. et Dupoux E. 2015, "The zero resource speech challenge 2015", In *INTERSPEECH-2015* : 3169-3173.
5. Ryzhova D., Kyuseva M. and Paperno. D. 2016, Typology of adjectives benchmark for compositional distributional models, in *Proceedings of the 10th Language Resources and Evaluation Conference* : 1253-1257.

Induction de grammaires, de propriétés typologiques et de lexiques

Documenter une langue, c'est aussi analyser sa syntaxe, produire une grammaire qui décrit ses contraintes structurelles et leurs interactions avec les contraintes lexicales.

Pour cette tâche, les techniques d'analyse syntaxique, d'induction de grammaire, de traduction automatique et d'alignement développées par la linguistique informatique apportent des perspectives nouvelles. Elles permettent notamment d'automatiser la détection des propriétés typologiques (place relative du nom et de l'adjectif ou ordre des constituants majeurs de la phrase par exemple) d'une langue[6] ou encore de créer une grammaire computationnelle qui pourra être testée sur les données disponibles[7]. Ces approches reposent généralement sur l'utilisation des gloses (1b) et des traductions (1c) que les linguistes de terrain associent aux données collectées (1a).

(1) a.	Jutta	khet	-a	-ŋ	-e
b.	Shoe	buy	PST	1sS	P-IND.PST
c.	I bought a pair of shoes				

L'alignement automatique est utilisé pour mettre en correspondance les éléments de la phrase à analyser avec ceux de sa traduction, l'analyse syntaxique pour associer cette traduction à un arbre syntaxique et la projection pour transposer l'analyse de la traduction. La Figure 1 illustre ce processus.

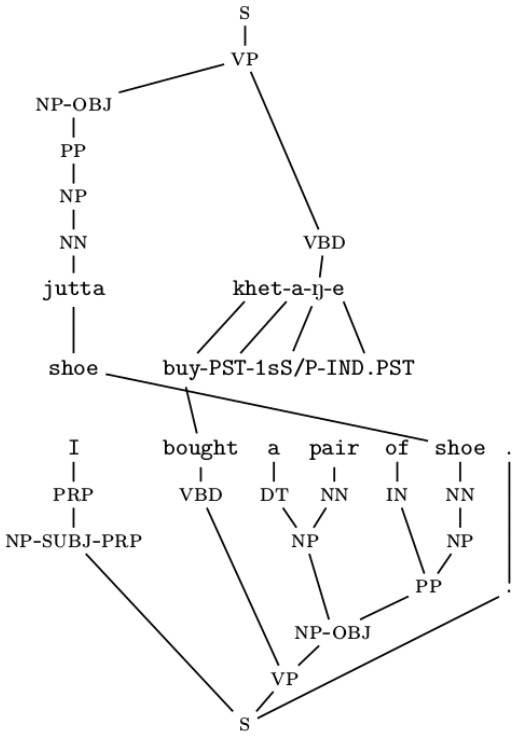


Figure 1 : Alignement et projection syntaxique d'une langue à l'autre

À partir de ces données enrichies, des propriétés typologiques peuvent être dérivées par des méthodes statistiques basées sur la fréquence d'apparition des phénomènes pertinents dans l'ensemble des arbres syntaxiques créés pour le corpus d'étude et des grammaires computationnelles peuvent être extraites grâce à une méthode existante qui permet de produire, à partir d'un ensemble de propriétés typologiques, la grammaire correspondante. Parce qu'elle peut être exploitée soit pour analyser des phrases soit pour en générer, cette grammaire informatique peut être validée et affinée à partir des données dont on dispose. Toutes les phrases sont elles couvertes par cette grammaire ? Les phrases générées par la grammaires sont elles grammaticales ? La grammaire informatique permet une confrontation systématique entre modèle et données. Elle facilite également la construction de corpus arborés et ainsi, l'analyse des interactions entre différents phénomènes.

Enfin, les techniques statistiques de traduction automatique peuvent être exploitées pour faciliter l'analyse des textes et la création de lexiques bilingues. Étant donné un corpus parallèle alignant des phrases et leur traduction, ces techniques permettent de trouver automatiquement la correspondance entre les expressions de deux langues et ainsi de créer rapidement et de façon semi-automatique, un proto-lexique listant les traductions des mots et des phrases courtes.

Conclusion

Les techniques de Traitement Automatique des Langues (TAL) ouvrent de nouvelles perspectives pour la documentation des langues en péril. Elles permettent d'accélérer la collecte des données, d'uniformiser les formats utilisés et d'automatiser la création de modèles informatiques (sons, unités lexicales, lexiques, grammaires, classes morphosyntaxiques, etc.) dont on peut vérifier, par le biais d'algorithmes d'analyse et de génération, la couverture et la précision.

Inversement, l'élargissement du TAL aux langues en péril favorise le développement de modèles informatiques plus adaptés à la description linguistique. En effet, les recherches dans ce domaine ciblent principalement des langues « bien dotées », c'est-à-dire, des langues pour lesquelles on dispose de larges corpus écrits. Pour adapter ces approches à des langues peu dotées, notamment les langues en voie d'extinction, de nouvelles méthodes d'apprentissage doivent être mises au point qui nécessitent moins de données, comme les méthodes basées sur le transfert, l'apprentissage actif ou l'apprentissage faiblement supervisé.

contact&info
▶ Claire Gardent
claire.gardent@loria.fr
Denis Paperno
denis.paperno@loria.fr
LORIA

6. Bender E., Goodman M., Crowgey J. and Xia F. 2013, Towards creating precision grammars from interlinear glossed text: Inferring large-scale typological properties, in *Proceedings of the 7th Workshop on Language Technology for Cultural Heritage, Social Sciences and Humanities*.
7. Bender E., Crowgey J., Goodman M. and Xia F. 2014, Learning Grammar Specifications from IGT: A case Study of Chintang, in *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*.