HAL
archives-ouvertes.fr

Favorisons la diversité linguistique en TAL

Chantal Enguehard, Mathieu Mangeot

▶ **To cite this version:**

Chantal Enguehard, Mathieu Mangeot. Favorisons la diversité linguistique en TAL. Journée d'étude de l'ATALA. "Ethique et Traitement Automatique des Langues", Nov 2014, Paris, France. 2014. <hal-01096592>

HAL Id: hal-01096592

https://hal.archives-ouvertes.fr/hal-01096592

Submitted on 17 Dec 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial - ShareAlike| 4.0 International License

Enguehard, C. Mangeot, M. Favorisons la diversité linguistique en TAL. Journée d'étude de l'ATALA. "Ethique et Traitement Automatique des Langues". 22 novembre 2014. Paris.

Favorisons la diversité linguistique en TAL

Chantal Enguehard[1], Mathieu Mangeot[2]
1 : LINA, BP 92208, F-44322 Nantes Cedex 03
2 : GETALP-LIG, BP 53 F-38042 Grenoble Cedex 9
chantal.enguehard@univ-nantes.fr, mathieu.mangeot@imag.fr

Mots-clé: Traitement Automatique des Langues, TAL, Traduction Automatique, TA, langues peu dotées, TIC pour le développement, diversité linguistique, ressources électroniques

1. Introduction

Dans le domaine du Traitement Automatique des Langues (TAL), la plupart des travaux portent sur des langues bien dotées qui ne manquent ni de ressources, ni d'outils. Les ressources en ligne peuvent être des lexiques, des dictionnaires, des corpus, etc. Il reste pourtant beaucoup à faire sur les langues peu dotées dont certaines ont un grand nombre de locuteurs (il y a, par exemple, 250 millions de bengalophones, 30 millions de haoussaphones).

Nous montrons dans cet article la faible diversité linguistique en TAL et évoquons quelques raisons pour cet état de fait. Nous pensons enfin que, sur la base de ce constat, la communauté pourrait contribuer à remédier à ce problème en adoptant quelques règles simples.

Dans la première partie, nous abordons la rareté et la faible qualité des ressources en ligne puis nous présentons la faible diversité linguistique des travaux en TAL à travers une étude sur les publications du domaine, ainsi que celle des dictionnaires et des systèmes de traduction automatique. Dans la deuxième partie, nous évoquons les besoins des locuteurs et des chercheurs en TAL. Dans la troisième partie, nous abordons certaines difficultés rencontrées lors de travaux sur les langues peu dotées. En conclusion, nous formulons des préconisations visant à augmenter la diversité linguistique du TAL.

2. Ressources et TAL

2.1. Ressources en ligne inaccessibles ou de faible qualité

Il existe des sites web présentant des ressources linguistiques pour des langues peu dotées mais peu permettent d'accéder effectivement à des ressources en ligne. Par exemple, glottolog[1], essentiellement dédié à la définition des langues (Nordhoff & Hammarström, 2011) inclut une collection de plus de deux cents mille références de travaux descriptifs de langue (grammaires, dictionnaires, lexiques, textes, etc.) mais ne donne pas accès à ces références.

Certains sites web issus, en partie, de travaux académiques présentent des données lexicales de langues peu dotées. En voici deux caractérisés par une volonté d'universalité.

– Le Rosetta project[2] a pour ambition de collecter et présenter une librairie numérique des langues. Un des projets est PanLex[3] qui recense des traductions de mots dans plusieurs milliers de langues (Kamholz et al. 2014) mais ne donne ni indication lexicale ou morphologique ni référence quant à l'origine de ces traductions. Sans faire une évaluation exhaustive, nous avons recherché les traductions de quelques mots dans des langues pour lesquelles nous possédons des ressources de bonne qualité. Voici des traductions affichées pour le mot anglais "*house*". Une traduction en bambara est "*bõ*" or la lettre *õ* ne fait partie de l'alphabet malien du bambara ; de même en haoussa, quatre des six traductions présentées font appraître des signes absents de l'alphabet nigérien. En kanouri, l'unique traduction est "*nǯiîm*" qui fait apparaître deux lettres qui n'appartiennent pas à l'alphabet kanouri. En français, cinquante traductions apparaissent mélangeant noms et verbes : nous relevons les étonnants "*gïte*" et "*députer*".

– le projet Kamusi[4] (Benjamin &. Radetzky 2014) est très récent. Bien qu'officiellement consacré à toutes les langues, il est pour l'heure essentiellement focalisé sur le swahili et l'anglais. On peut regretter que le site ne référence pas le système d'écriture utilisé pour chaque langue ainsi que les origines des mots, définition et liens de traduction présentés.

Cet examen nous a permis de constater les limites des sites participatifs. Des graphies fantaisistes apparaissent et il n'est pas certain, en ce qui concerne les langues ayant peu de locuteurs alphabétisés, que ces erreurs soient un jour corrigées. Du fait du manque de documentation, de tels sites ne peuvent être considérées comme des ressources fiables, quelles que que soient les utilisations (simple consultation ou TAL)

Il existe aussi pléthores de sites web amateurs ou à vocation commerciale présentant des lexiques, des dictionnaires ou encore des textes. Mais ces ressources ne sont généralement pas documentées, il est fréquent que les auteurs ne soient pas identifiés. Les missionnaires du Summer Institute of Linguistics (S.I.L.) constituent un cas

1 http://glottolog.org/

2 http://rosettaproject.org/
3 http://www.panlex.org/
4 www.kamusi.org

particulier dans la mesure où cet institut contribue à produire et distribuer de grandes quantités de données linguistiques de basse qualité. Par exemple, il est possible d'accéder via son site web à une page pointant vers plus 300 dictionnaires. Les dictionnaires haoussa et estonien que nous avons examinés apparaissent inutilisables tant leur qualité est faible.

2.2. Diversité linguistique des travaux en TAL

Nous avons exploré 7200 articles de recherche publiés dans l'une des deux conférences ACL ou LREC entre 2000 et 2014 afin de rechercher la mention de langues (parmi une liste de 209 langues) dans le titre, le résumé ou le corps de l'article (sans les références).
Globalement, environ 90% des articles de LREC et 75% de ceux d'ACL mentionnent au moins une langue.
Pour les deux conférences, l'anglais est mentionné dans plus de 60% des articles. Le français et l'allemand apparaissent ensuite dans des proportions très similaires. Pour ces trois langues, cette proportion tend à augmenter avec le temps pour les conférences LREC.

	anglais	alle-mand	fran-çais	chinois	espa-gnol	japo-nais
ACL & LREC	64%	21%	20%	17%	15%	12%
ACL	62%	14%	12%	23%	9%	12%
LREC	65%	25%	25%	13%	19%	12%

Table 1 : Les six langues très bien dotées (mentionnées dans plus de 10% des articles)

Ce classement fait ensuite apparaître des langues indo-européennes au milieu desquelles s'intercalent le japonais et l'arabe, puis le coréen et l'hindi. Les dix langues les plus mentionnées le sont dans 79% des articles. *A contrario*, les cent cinquante langues les moins citées n'apparaissent que dans 5% des articles. En voici quelques-unes : le kazakh, le ouïghour, le wolof, le groënlandais, le kashmiri.
La recherche de couples de langues[5] a extrait 78 langues apparaissant au sein de 449 couples. 80% des occurrences font apparaître l'anglais. Cette proportion est moins forte pour les articles de LREC (72%) par rapport à ACL (91%) sans évolution temporelle notable.
L'anglais est apparié à 65 langues. Suivent le français, l'allemand et l'espagnol appariés avec 20 à 30 autres langues. Le couple anglais-chinois est apparu le plus fréquemment. Le premier couple ne mentionnant pas l'anglais est le français-allemand (en 17e position) suivi du chinois-japonais.
La centralité de l'anglais et, dans une moindre mesure, des langues très bien dotées, apparaît nettement.

5 Nous n'avons pas distingué les couples de langues selon leur orientation : langue source - langue cible. Par exemple, anglais-français et français-anglais sont ici considérés comme mentionnant le couple de ces deux langues.

2.3. Dictionnaires bilingues

Le dictionnaire (monolingue ou bilingue) est souvent considéré à juste titre comme une ressource de base et un élément essentiel de tout système de TAL. Nous évoquons ici deux contextes différents de dictionnaires bilingues français-autre langue : le khmer et le japonais.

2.3.1. Contexte français-khmer

Nous avons étudié ce contexte dans le cadre d'un projet d'informatisation du dictionnaire français-khmer de Denis Richer et l'association Pays Perdu (Mangeot, 2014). Le khmer est une langue peu dotée (Berment, 2004). Il n'est donc pas étonnant de trouver peu de ressources disponibles en ligne concernant cette langue. Le Cambodge étant une ancienne colonie française, nous aurions pu penser qu'il existait au moins un dictionnaire français-khmer en ligne. Or, le seul dictionnaire disponible est un dictionnaire anglais-khmer[6]. Les khmérophones voulant comprendre le français sont alors obligés de passer par l'anglais. Par ailleurs, nous n'avons trouvé que deux mentions de cette langue dans notre corpus d'articles de recherche (voir 2.2).

2.3.2. Contexte français-japonais

Depuis plusieurs années nous suivons avec attention l'évolution du contexte français-japonais pour notre recherche (Mangeot et al., 2003) mais aussi pour notre utilisation personnelle.
Le français et le japonais sont, sans conteste, deux langues très bien dotées. Il n'en va pourtant pas de même pour le couple de langues français-japonais. Il n'existe pour l'heure pas encore de dictionnaire français-japonais d'une couverture correcte (supérieure à 20 000 mots) qui soit consul-table en ligne. Comme nous le démontrerons par la suite, il n'existe pas non plus de système de traduction directe français-japonais disponible en ligne ou à l'achat.
Il existe bien sûr des dictionnaires français-japonais de bonne qualité en version électronique mais ils restent la propriété des maisons d'édition qui n'osent même pas les mettre à disposition sur le Web. Pour y accéder, il faut acheter au prix fort (200 € environ) un dictionnaire électronique de poche.
Il existe pourtant plusieurs dictionnaire anglais-japonais de bonne couverture, disponibles en ligne et au téléchargement comme par exemple le très connu JMDict[7] bien plus complet en anglais que dans les autres langues. En revanche, dans notre recensement, le couple français-japonais apparaît effectivement rarement. Finalement, et comme pour le khmer, les francophones étudiant le japonais utilisent ce dictionnaire, ce qui les oblige à maîtriser l'anglais ou à utiliser un autre dictionnaire français-anglais avec, donc, des risques accrus de contresens ou d'approximations.

6 http://www.english-khmer.com
7 http://www.edrdg.org/jmdict/j_jmdict.html

2.4. Couples de langues en traduction automatique (TA)

Depuis l'avènement de la traduction automatique statistique, les systèmes de TA se sont multipliés. Ils restent cependant tributaires de l'existence de corpus bilingues. Ceux-ci ne sont pas faciles à trouver pour de nombreux couples de langues. Ils sont souvent constitués à partir de ce que nous appelons pivots anglais indirects : des livres anglais traduits dans de nombreuses langues (par exemple la série des Harry Potter) ou sous-titres de films (qui sont pour majorité des films anglophones).
Nous relatons ici deux systèmes de traduction automatique : celui de la société Systran qui dispose de systèmes hybrides et celui de Google fondé uniquement sur des statistiques.

2.4.1. Le système SystranSoft

Systran Enterprise Server 7[8] comporte 52 paires de langues hybrides dont 15 pour l'anglais, 7 pour le français, 5 pour l'allemand, 4 pour l'espagnol, l'italien et le portugais, 28 paires de langue additionnelles dont 24 pour l'anglais et 4 pour le français.
Le français est bien représenté (Systran est une entreprise française). Si l'on se réfère au tableau du nombre de locuteurs natifs d'une langue de Wikipédia[9], on remarquera que dans ces couples de langues, on ne trouve pas le bengali (7ème langue), le javanais (10ème) ou le pendjabi (11ème).

2.4.2. Google Translate

Google Translate[10] propose actuellement 80 langues disponibles. Il est possible de traduire de n'importe quelle langue de cette liste vers n'importe quelle autre. 3 160 couples de langues qui sont donc proposés. Ce qui implique bien sûr pour la plupart de ces couples de passer par une langue pivot (l'anglais) mais Google ne mentionne pas lorsqu'un pivot est utilisé. Il est possible de s'en apercevoir au détour d'une traduction lorsqu'un mot anglais est affiché dans le résultat ou qu'un résultat n'a aucun rapport avec le texte source alors que ni la source, ni la cible ne sont l'anglais. Par exemple, la traduction japonaise des expressions « *La vie n'est pas un long fleuve tranquille* » ou « *C'est la cerise sur le gâteau* » donnent une traduction mot à mot des expressions anglaises équivalentes et qui n'ont aucun sens en japonais : « *jinsei wa bara no betto dewa arimasen* » [la vie n'est pas un lit de roses] (Life is not a bed of roses) et « *Kore wa kēki no ue no aishingu desu* » [C'est le « aishingu » sur le gâteau] (This is icing on the cake).

8 http://www.systransoft.com/translation-products/server/systran-enterprise-server/language-pairs/
9 http://fr.wikipedia.org/wiki/Liste_des_langues_par_nombre_de_locuteurs_natifs
10 http://translate.google.com/about/intl/fr_ALL/

3. Besoins

3.1. Besoins humains

Être locuteur d'une langue peu dotée se concrétise par un accès limité ou nul aux ressources linguistiques de cette langue : les dictionnaires, mais aussi les manuels scolaires, les œuvres littéraires, la presse, etc. Ces besoins insatisfaits concernent de nombreux aspects de la vie : web, éducation, santé, culture, etc. (Osborn, 2011).
Au-delà du seul accès aux ressources, c'est la possibilité de s'exprimer, c'est-à-dire d'exercer sa liberté d'expression qui est en jeu.
L'UNESCO a évoqué à plusieurs reprises cette dimension ainsi que la richesse que constitue la diversité linguistique :

> « *les États membres de l'UNESCO, en décidant de célébrer les langues maternelles, ont voulu rappeler qu'elles constituent non seulement un élément essentiel du patrimoine culturel de l'humanité mais aussi l'expression irréductible de la créativité humaine dans toute sa diversité* » Paris, 21 février 2000.
>
> « *L'UNESCO, en encourageant la construction des sociétés du savoir plurielles et inclusives, reconnaît que la langue représente un facteur crucial dans l'aptitude à communiquer. La capacité des peuples à partager et à accéder au savoir, à posséder des moyens d'action dans les sociétés du savoir et à pouvoir participer au monde numérique dépendra de plus en plus des solutions multilingues à leur disposition*.» Bamako, 6 et 7 mai 2005.

3.2. Besoins scientifiques

Les besoins relèvent également du domaine scientifique : les milliers de langues peu dotées sont autant de *terra incognita* qui restent à explorer, la première étape étant la constitution de ressources électroniques. Les défis scientifiques sont donc légion.
Certaines problématiques concernent des champs qui ont été investis par des organisations non académiques. C'est le cas de la question non triviale de l'identification des langues : le standard ISO 639-3, fondé sur les travaux de la SIL, et utilisé pour coder les noms de langues, s'avère inadapté du fait de l'absence de définition claire des langues qu'il nomme (Nordhoff & Hammarström, 2011).

4. Écueils

4.1. Difficultés à mener des travaux de recherche

Comme les populations, les recherches en TAL sont confrontées à la pénurie de ressources linguistiques de bonne qualité ce qui diminue drastiquement le champ des possibles. Les recherches sont à la fois sporadiques, disséminées et volatiles (Streiter et al., 2006).
Le contexte socio-économique des langues peu dotées est caractérisé par des ressources réduites : il y a peu de linguistes ayant une langue peu dotée comme langue maternelle et exerçant leur activité professionnelle sur cette langue, et le budget consacré au développement de ressources linguistiques est faible. Améliorer la création et

l'accès à des ressources linguistiques constitue donc un problème difficile.
L'enjeu est considérable car la création de ressources pérennes et de qualité constitue un goulet d'étranglement qui engendre un cercle vicieux :l'absence de ressources décourage les recherches en TAL, celles-ci font défaut dans le développement et l'épanouissement des populations, y compris en termes d'éducation, ce qui entraîne un faible niveau d'études et une pénurie de « travailleurs de la langues » que sont les linguistes, lexicologues, écrivains, etc.
Financer des projets de développement de ressources pour les langues peu dotées reste hasardeux : les ressources sont rares et limitées. Quelques possibilités subsistent pourtant comme le montrent les projets récents auxquels nous participons : le projet DiLAF[11] d'informatisation de dictionnaires langues africaines-français a été financé par le fond francophone des inforoutes et le projet ALFFA[12] (African Languages in the Field - Fundamentals and Automation) est financé par l'Agence Nationale pour la Recherche (sans financement des partenaires non français).

4.2. Difficultés à publier

Paradoxalement, alors que les publications sur les langues peu dotées apportent beaucoup d'informations du fait de leur rareté, et que la production de ressources électroniques est cruciale (autant pour les populations concernées que pour le TAL), elles sont souvent rejetées en raison de leur manque de nouveauté quant à leur apport purement théorique.
Il faut pourtant déployer des trésors d'imagination pour développer des ressources et outils dans des contextes peu favorables. En effet, comment développer un système de traduction automatique statistique lorsqu'il n'y a pas ou très peu de corpus disponible ? Comment faire travailler une équipe de contributeurs lorsqu'il n'y a pas d'accès à Internet et que les coupures d'électricité sont quotidiennes ? La nouveauté tient alors dans le développement de méthodologies innovantes, à la fois simples, économiques et reproductibles, ce qui inclut le transfert de connaissances et de bonnes pratiques. Il s'agit donc d'intégrer cette dimension organisationnelle et humaine dans l'évaluation des recherches au lieu de simplement l'écarter.
Une autre difficulté plus grande encore est celle que rencontrent nos collègues du Sud pour publier dans des conférences du domaine. Un chercheur CNRS parisien et un enseignant-chercheur nigérien sont loin d'avoir les mêmes conditions de travail. En particulier le chercheur africain doit prendre en compte des limites (financières, d'accès à l'information, d'accès à Internet, isolement scientifique, etc.). Mais lorsqu'il s'agit d'évaluer une publication, le contexte du travail et les limites ne sont pas pris en compte. Les évaluations de ces publications sont peu adaptées car les difficultés sont souvent méconnues et les stratégies pour les résoudre ne sont pas reconnues.
Des alternatives existent mais elle sont rares, comme celle de publier dans le domaine des TIC pour le développement[13].

5. Conclusion

Malgré leur apport indéniable, tant scientifique que sur le plan du développement humain, et donc de l'éthique, les travaux en TAL sur les langues peu dotées rencontrent de nombreuses difficultés. Celles-ci ne sont pas insurmontables et la communauté pourrait s'emparer de ce problème et adopter certaines mesures destinées à favoriser ces travaux.
Concernant les travaux réalisés sur ces langues, il est nécessaire d'organiser le partage des données linguistiques, des méthodologies et des outils. La communauté anglophone a déjà œuvré dans ce sens avec le site web africanlanguages.org (qui a malheureusement dû cesser son activité suite au manque de bénévoles, ce qui montre, de nouveau, la faible pérennisation des travaux de ce domaine). Certaines données ont heureusement été reprises sur le site de l'aflat.org.
Concernant les publications, nous suggérons un aménagement des formats des articles de recherche en TAL. Il s'agirait d'ajouter une rubrique mentionnant explicitement les langues (ou familles de langues) prises en compte par la recherche exposée. Cette mention permettrait de faire un suivi plus précis des langues traitées. Elle pourrait être aussi l'occasion d'un questionnement chez les chercheurs, et ainsi encourager la diversification des langues traitées.
L'évaluation des publications pourrait également adopter de nouveaux critères prenant en compte le contexte de travail des collègues (pays du Nord/ du Sud) ainsi que leur objet d'étude (langue bien/peu dotée).

6. Bibliographie

Benjamin, M &. Radetzky, P. Small Languages, Big Data: Multilingual Computational Tools and Techniques for the Lexicography of Endangered Language. (2014). 14th Conference of the European Chapter of the Association for Computational Linguistics. Gothenburg, Sweden, 26-30 April 2014.

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

Kamholz, D. Pool, P. Colowick, S. (2014). PanLex: Building a Resource for Panlingual Lexical Translation. LREC 2014.

Mangeot, M. (2014). *MotàMot project: conversion of a French-Khmer published dictionary for building a*

11 http://www.dilaf.org/Home.po
12 http://getalp.imag.fr/xwiki/bin/view/Projects/ALFFA
13 Voir par exemple la revue Information Technology for Development.

multilingual lexical system. Proc. of LREC 2014, Reykjavik, Island, 28-30 May 2014, 8 p.

Mangeot, M. & Enguehard, Ch. (2013). Des dictionnaires éditoriaux aux représentations XML standardisées. Chapitre 8. In N. Gala, M. Zock (Eds.), *Ressources Lexicales : contenu, construction, utilisation, évaluation.* Linguisticae Investigationes Supplementa, John Benjamins Publishing, Amsterdam, Pays-Bas, 24 p.

Mangeot, M. Sérasset, G. & Lafourcade, M. (2003). *Construction collaborative de données lexicales multilingues, le projet Papillon.* Revue TAL, édition spéciale, Les dictionnaires électroniques : pour les personnes, les machines ou pour les deux ? M. Zock, J. Caroll (Eds.) Vol. 44:2/2003, pp. 151-176.

Nordhoff S. & Hammarström, H. (2011). Glottolog/langdoc: Defining dialects, languages, and language families as collections of resources. In Proceedings of the 10th International Semantic Web Conference (ISWC 2011).

Osborn, D. (2011). *Les langues africaines à l'ère du numérique.* Laval, Canada: Presses de l'Université Laval.

Streiter O., Scannell K., Stuflesser M. (2006). *Implementing NLP projects for non-central languages : Instructions for funding bodies, strategies for developers.* Machine Translation, volume 20.