Constitution semi –automatique de bases onto-terminologiques à partir de corpus numérisés: comment repérer et combler les lacunes dans les arborescences terminologiques des langues à faible diffusion

Par **Violaine PRINCE**
Université Montpellier 2 et LIRMM-CNRS

Introduction

Les langues à faible diffusion, c'est-à-dire des langues parlées par des communautés restreintes en nombre, ou peu écrites, peuvent aujourd'hui bénéficier des bienfaits de la numérisation au lieu d'en être les victimes.
En effet, si la numérisation et la diffusion électronique, via Internet, tendent à favoriser les supports langagiers les plus populaires, elles sont aussi la voie de secours des langues faiblement diffusées, car elles peuvent assurer leur pérennité, sous forme de mémoire, et de conservation patrimoniale (V.B.BACHIMONT, 2013).La notion de "document" devient alors l'enjeu principal et l'acteur essentiel de cette pérennité (V.M.IHADJADEN et al. 2010). Ce document nécessitera une écriture, directement produite ou bien issue d'une transcription.

Elles peuvent aussi promouvoir la diffusion des langues à travers des blogs, fora ou réseaux sociaux, mais également sous forme didactique (V. S.DETEY et al. 2010). Dans ce cas, l'aspect écrit peut être partiellement relayé par d'autre medias (e.g. images avec légendes, vidéos, bandes sonores).

Il est donc essentiel aujourd'hui d'examiner ces deux aspects positifs de la numérisation, afin d'en doter plus largement les langues dites "minorisées".

Notre communication s'adresse plus particulièrement à la première branche, c'est-à-dire à l'enrichissement de la mémoire et des ressources numériques d'une langue. Elle se penchera plus en avant sur une méthodologie de construction de ressources lexicales, c'est-à-dire d'objets linguistiques ayant pour but de présenter et/ou d'expliquer des mots ou des termes de cette langue.

Si l'archivage et le stockage documentaire sont une fonction importante d'un point de vue de la documentation, les ressources lexicales sont l'étape première dans la construction patrimoniale numérique qui aurait pour effet de faire vivre la langue et d'en augmenter la diffusion, pas simplement d'en conserver les produits. En effet, les dictionnaires issus de la recherche en traitement automatique des langues, en lexicométrie, ou en linguistique quantitative, n'ont jamais été aussi nombreux. Ces dictionnaires se déclinent en deux grandes catégories, elles-mêmes divisées en deux sous catégories :

1. Des dictionnaires de type linguistique, monolingues (1) ou bilingues (2), obtenus à partir de corpus numérisés monolingues, dont on traitera brièvement de la création dans le premier paragraphe de cette intervention,
2. Des structures de nature plus propre à être exploitées automatiquement, ou semi-automatiquement, que l'on nommera bases onto-terminologiques (en expliquant l'origine de cette appellation) et dont l'objectif est de produire des ressources directement réexploitables pour assurer la pérennité de l'usage de la langue dans l'univers numérique : on distinguera alors les onto-terminologies de spécialité (langages techniques) (1) de celles qui sont plus proches des dictionnaires de langue (ontologies de la langue telles que Wordnet) (2).

Dans cette communication, nous nous intéresserons plus particulièrement aux onto-terminologies, leur importance pour les langues à faible diffusion, et nous proposerons une méthodologie susceptible d'aider à leur construction. Nous ferons un bref tour d'horizon de

l'apport des techniques d'origine informatique pour l'enrichissement des ressources patrimoniales des langues, en nous intéressant d'abord à la création de dictionnaires (linguistiques) à partir de corpus, pour soutenir l'action lexicographique et lexicologique. Puis nous nous concentrerons sur la création de ressources onto-terminologiques, en nous focalisant plus particulièrement sur les chaînons manquants, c'est-à-dire ces termes qui manquent dans l'univers des langues à faible diffusion, pour les transformer en langues "complètes" et compétitives sur le plan de la couverture linguistique, étape primordiale pour soutenir leur diffusion.

I- Les ressources lexicales: constitution de dictionnaires à partir de corpus numérisés

L'existence d'une langue dans le carousel académique tient souvent à peu de chose, et le nombre de dictionnaires, monolingues ou bilingues qui traitent de ses objets langagiers, semble être un élément passablement significatif du rayonnement de cette langue. Lexicologues, lexicographes, courageux constructeurs de dictionnaires ont par le passé fourni un travail acharné nécessitant de longues années avant de produire des ressources qui ont contribué à asseoir les fondements académiques de ces langues.

Ainsi, nombre d'entre elles ont été sauvées de l'extinction en raison de l'acharnement désintéressé de lexicologues, et d'autres communications dans ce colloque pourront fournir des exemples frappants de tels travaux d'érudition nécessitant l'intervention de plusieurs disciplines scientifiques telles que l'anthropologie, la linguistique, l'ethnologie, la géographie, etc.
Un des verrous les plus importants parmi ceux auxquels s'était heurté, historiquement, le traitement automatique des langues (TAL), était lié à l'existence et la constitution de corpus numériquement exploitables, à partir desquels les techniques mises au point par le TAL sont en mesure d'extraire des termes pour fabriquer des dictionnaires en partant du discours. Le problème a toujours été de disposer de données informatisées en nombre suffisant pour produire des ressources.

Ce verrou a été en grande partie levé, dans les années 1990, grâce à la très large diffusion sur le Web de documents numériques qui se sont comportés en corpus. La numérisation du patrimoine documentaire, dont nous avons parlé en introduction, mais surtout, la diffusion sans restriction de documents numériques, a permis un essor sans précédent de la fouille de textes, domaine du TAL où les données textuelles sont fournies en grande quantité, et à partir desquelles des tâches classiques de l'informatique documentaire, telles que l'indexation ou la recherche d'information (V.G.SALTON et al. 1975) ont retrouvé un éclat qui fut éclipsé durant les deux précédentes décennies par des techniques plus "cognitives". Aujourd'hui, la recherche d'information atteint des degrés de popularité sans précédent grâce aux moteurs de recherche (e.g. Google, Yahoo, etc.) qui font partie du quotidien des usagers des dispositifs de communication les plus répandus. Beaucoup de travaux ont produit des ressources lexicales à partir de corpus, en particulier des dictionnaires. On distinguera trois types de productions :

1. Les dictionnaires monolingues (langues faiblement dotées)
2. Les dictionnaires bilingues, ou pluri-lingues
3. les dictionnaires explicatifs, de synonymes, etc.

1.1 Dictionnaires monolingues pour langues faiblement dotées

1.1.1 Création *ab initio*

Il s'agit de langues possédant peu de ressources, de langues mortes ou mal décrites. Il faut alors disposer de corpus écrits, soit par transcription de corpus oraux, soit déjà présents mais dont il faut réaliser la numérisation (e.g, égyptien ancien chez (Rosmorduc, 2002). La création des dictionnaires se fait par association entre un terme t, et ses voisins, lorsqu'une structure syntaxique telle que :

t est x y z...apparaît.

Le verbe être peut être remplacé par d'autres verbes d'explication ou de description (apparaître, sembler, se décrire, se composer, etc...). Cette structuration en triplets (terme, verbe de description, (liste de descripteurs)) est un patron qui a par exemple fait ses preuves dans le peuplement d'ontologies techniques (V.J. MAKKI et al. 2008)) et rien n'empêche de la considérer comme une base de départ, lorsque l'on dispose de très peu d'outils.

Il est clair que ces dictionnaires peuvent être fortement entachés d'erreurs, et seule une confrontation avec des natifs (si la langue est toujours parlée) peut permettre de corriger les attributions.

> *L'hypothèse fondamentale étant celle de Harris*[181] *, seule une grande masse de données ou une présence experte peuvent contrebalancer les erreurs des premiers résultats.*

C'est pourquoi, on préfèrera utiliser cette technique pour augmenter des dictionnaires déjà présents mais incomplets.

1.1.2 Amélioration de dictionnaires existants

Les langues faiblement dotées ne sont pas toujours complètement dépourvues, et il existe souvent des ébauches de dictionnaires, ou bien des dictionnaires "datés" et donc incomplets.

L'idée alors consistera à utiliser les corpus numériques cette fois-ci pour enrichir ces ressources et/ou les actualiser.

Plusieurs techniques peuvent être utilisées, selon que l'on possède ou pas des ressources appropriées.

1. Des dictionnaires bilingues entre la langue L (objet d'étude) et une langue L', mieux dotée,
2. Des corpus dans la langue L'
3. Des corpus dans la langue L
4. Des dictionnaires monolingues en L'

La combinaison (1, 2, 3,4) étant la plus complète, elle permet d'ajouter au dictionnaire de L les éléments suivants :

1. les termes *t* de *L*, apparaissant dans un corpus de *L* (3) qui existent dans le dictionnaire bilingue, (1) mais pas dans le dictionnaire monolingue
2. Si les corpus (2) et (3) sont comparables (alignables partiellement ou totalement, que ce soit par des méthodes automatiques ou semi-automatiques), tous les termes *t* de *L* qui sont alignés avec des termes *t'* de *L'*, qui existent ou non dans le dictionnaire bilingue (2), mais où les termes t' existent dans le dictionnaire monolingue de *L'* (4), ou à défaut, les termes *t'* de *L'* sont ajoutés au vocabulaire de *L*.

Cette solution permet d'intégrer des néologismes consacrés, des emprunts, des termes nouveaux. C'est ainsi que des termes tels que "computer", ou des acronymes tels que "DVD" peuvent entrer dans des langues faiblement dotées par le biais de leur existence dans des corpus de la langue faiblement dotée par emprunt, et de leur consécration dans une autre langue.

L'analyse automatique de ces corpus ne nécessite pas beaucoup d'outils de TAL. Un séparateur de termes peut être suffisant (classe "word" dans Java) si la langue *L* est peu fléchie. En revanche, si *L'* est mieux dotée, l'existence d'un étiqueteur morphosyntaxique ("tagger") et de procédures d'alignement (pour les corpus (2) et (3)) (V. Segura et Prince 2011) sont particulièrement bienvenues.

[181] ZELLIG Harris (1954) considère que le sens d'un terme est essentiellement défini par la distribution de ses voisins, dans toutes les occurrences où il apparaît

1.2 Dictionnaires multilingues pour langues faiblement dotées

Les langues faiblement dotées peuvent voir augmenter leurs chances de diffusion si on augmente le nombre de dictionnaires multilingues entre ces langues et des langues mieux diffusées. Ainsi, des langues telles que le coréen, le finnois, l'estonien, qui sont moins populaires que le japonais, le russe, le chinois ou l'anglais, ont largement bénéficié de la numérisation et des techniques de TAL pour augmenter le nombre de leurs ressources.

Les techniques de création ou d'augmentation de dictionnaires multilingues nécessitent les mêmes séries de ressources que l'augmentation des dictionnaires monolingues.
On notera cependant que l'alignement des corpus, ou l'existence de corpus parallèles ou comparables devient beaucoup plus prépondérante si on ne veut pas utiliser des systèmes de représentation pivot telles que UNL (Unified Natural language) (V. G.SERASSET et C.BOITET 1999). On peut cependant utiliser des langues comme pivot si on ne possède pas assez de ressources directes, ou si on n'a pas de corpus comparables. Par exemple, si on possède un dictionnaire français-anglais, et un dictionnaire anglais-japonais, on peut contribuer à la création d'un dictionnaire français-japonais en stoppant la dérive des fausses équivalences (la relation de traduction n'étant malheureusement pas transitive) grâce à des techniques statistiques sur les coccurrences de termes dans de grands corpus de la langue japonaise (V. K.TANAKA et V. PRINCE ,1995).

Cette technique sert également à la création de dictionnaires de spécialité (V.S. FERRARI et V. PRINCE ,1996), très utiles pour les langues à faible diffusion, puisque c'est souvent dans les domaines techniques qu'elles présentent des lacunes.

1.3 Dictionnaires explicatifs, de synonymes

C'est dans ce domaine que l'on trouvera pas mal de travaux numérisés, avec toutefois une confusion entre des dictionnaires "ontologies" (comme Wordnet, l'atlas sémantique de Ploux Ploux et al. 2010), etc.) et des dictionnaires linguistiques. Ces derniers sont produits à partir de corpus, et de travaux lexicographiques. Citons par exemple l'initiative du Projet Gutenberg qui a eu, entre autres numérisations, le souci de numériser le Roget (dictionnaire de synonymes de la langue anglaise, dont la structure peut être transformée en ontologie), le dictionnaire de synonymes de l'Université de Caen (V.S.PLOUX et B.VICTORRI, 1998), et les versions en ligne des dictionnaires de langue (pour le français, Hachette des synonymes, Grand Robert, Thésaurus Larousse, etc.).

II Bases onto-terminologiques : l'avenir pour l'augmentation de la diffusion des langues faiblement diffusées :

Il existe une différence entre un dictionnaire dit "linguistique" et une base onto-terminologique. Le premier est un texte dont la structure est assez "faible», de type linéaire, et qui suit le modèle :
ENTREE LEXICALE [numéro] [(flexion)] (catégorie [attribut]) : {texte d'explication.
Exemple}. Où les éléments entre [] sont facultatifs, et les éléments entre {} sont répétables.
Par exemple, on trouvera :
FRÉTILLANT (E) : (adj.) qui frétille, est animé de mouvements vifs : des poissons frétillants.

L'exploitation informatique directe d'un tel dictionnaire est difficile. Son enrichissement également : il nécessite une large réécriture, si ce n'est de la partie gauche, du moins de toute la partie explicative et illustrative. C'est pourquoi, on s'est très tôt penché sur des structures plus élaborées, relevant du modèle des graphes, dont l'algorithmique est un des domaines les plus répandus dans la recherche en informatique. L'utilisation de graphes comme mode de représentation des connaissances lexicales a donné lieu à des "ontologies" de la langue. Ces structures sont très utiles dans le Web sémantique, les applications informatiques, le TAL, la recherche d'information, l'indexation, etc. Et donc tout ce qui peut apporter de la diffusion pour une langue donnée.

2.1 Ontologies et Onto--terminologies

On ne redéfinira pas ici le terme "ontologie", que l'informatique a fortement galvaudé par rapport à sa définition originelle (description des propriétés de l'être, au sens philosophique de ce terme). On acceptera, pour des besoins pédagogiques, une description opérationnelle de ce mot sous forme de : *toute représentation des relations régulières entre concepts sous formes d'arêtes de graphes, les sommets étant étiquetés par les concepts.*

De manière traditionnelle, on restreint les relations régulières aux relations suivantes :

- relation d'inclusion : X est une sous-classe de Y (X et Y étant des classes d'équivalence sur certaines propriétés), ou X est une partie de Y (sans en être une sous-classe)
- relation d'appartenance : x est un élément de la classe X (également appelé instance ou occurrence)

Exemples :

Les cations sont des ions (Cation sous-classe de Ion)
Les mains sont une partie du corps (relation "partie-de")
Félix est un chat (élément de la classe des "chats").

Ces relations étant hiérachiques, les parties connexes de ce graphe sont des arbres. On peut créer par exemple l'arborescence suivante de racine "ion" : cation < ion de la relation d'inclusion de classe (relation sorte-de), mais on peut créer aussi l'arbre suivant : proton ¡ atome ¡ molécule ¡ substance de la relation "partie-de", avec comme racine "substance". Les relations sont transitives et asymétriques, elles sont donc considérées comme des relations d'ordre (l'aspect strict de l'ordre peut être inclus ou non).
La relation d'appartenance peut être vue comme une relation d'inclusion de classe entre un élément et une classe, et donc comme un cas particulier de cette relation.

Les sommets de ces arbres sont étiquetés à l'aide de "concepts". Ce terme est assez vague dans la littérature idoine et semble désigner un symbole qui n'est pas obligatoirement un mot ni un groupe de mots. La collusion entre un concept et des réalisations langagières relève plus du problème de l'ontologie lexicale, c'est -à -dire de la structuration graphique des mots de la langue, qui est le cas qui nous intéresse.

Dans cette vision, l'hyperonymie-hyponymie peut être vue comme une matérialisation de la relation d'inclusion de classe, et la méronymie -holonymie comme celle de la relation "partie-de".
Il est clair que le modèle ontologique hiérarchique ainsi défini est beaucoup trop pauvre par rapport aux besoins de la langue et même par rapport à l'organisation des connaissances, qu'elles soient lexicalisées ou non. D'autres relations ont été rajoutées, comme la relation d'attribution permettant de revenir vers des modèles de réseaux sémantiques et donc de perdre l'aspect hiérarchique et arborescent. Le modèle de Wordnet (V.G.A. MILLER, 1995) recense également des relations de synonymie et d'antonymie. Le problème est que ces dernières ne sont pas fiables, en raison de la polysémie-homonymie (multiplicité des sens d'une chaîne de caractères).

Un modèle tel que celui de Jeux de Mots (V. A.JOUBERT et M. LAFOURCADE, 2008) recense une trentaine de types différents de relations.

Les onto-terminologies sont introduites lorsque, justement on veut représenter des concepts, en général d'un domaine particulier, et que l'on se heurte à l'emploi des mots dans ce domaine.

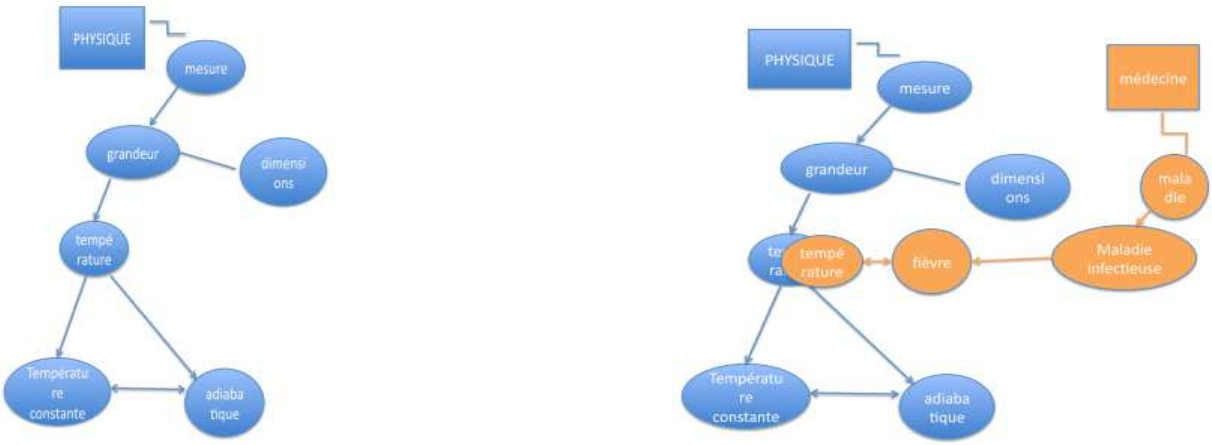
On va donc rajouter une relation dite "terminologique", qui signifie qu'un objet x est un terme désignant le concept y, dans le contexte du Domaine D.

Par exemple, l'utilisation du terme "température" dans le domaine de la physique, va et lié à un concept lexicalisé par "grandeur", qui signifie "mesure", et également à des termes tels que "mouvement", "agitation" etc.

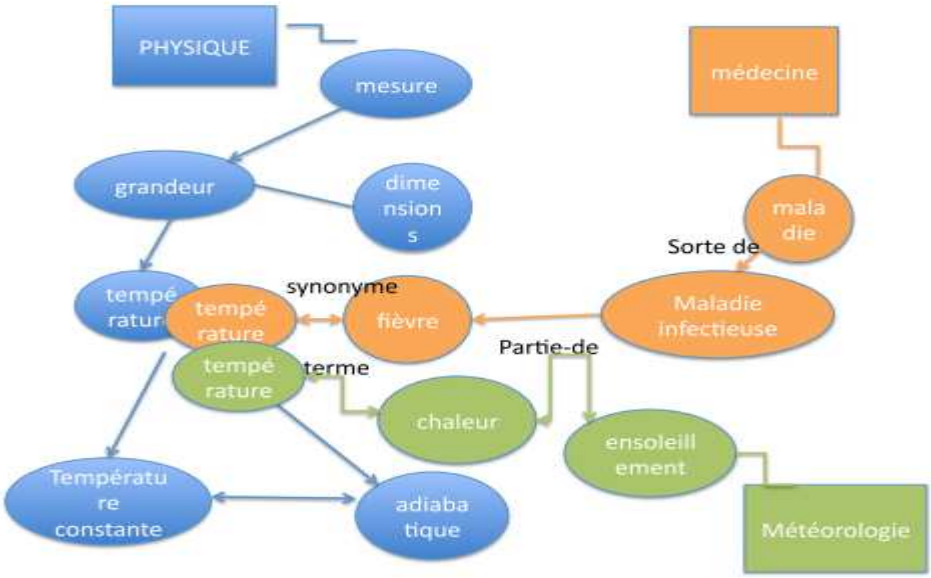
Le même terme en médecine, va être lié au concept lexicalisé par "maladie", ce qui n'est pas le cas de la physique.
Enfin, dans le domaine de la météorologie, il va être lié à ceux d'"ensoleillement", de "dépression", d'"anticyclone", etc.

On voit donc que la terminologie réalise plusieurs fonctions :

1. Elle fournit des formes lexicales à des sommets de graphes qui entretiennent des relations typées avec d'autres sommets, ces relations étant matérialisées par des arêtes de graphe. Ces formes lexicales peuvent être composées : par exemple le terme composé "élévation de température" existera en médecine, et sera lié à celui de "fièvre", alors qu'"élévation des températures" (forme plurielle pour le deuxième nom) existera en météorologie et sera lié à celui de "chaleur".
2. Elle permet de spécialiser les voisins d'un sommet en fonction d'un domaine : leurs attributs, leurs relations ne sont plus les mêmes
3. Elle permet de prédire les possibles utilisations d'un terme dans un domaine en fonction de ses voisins et des relations qu'il entretient avec eux (V. A.JOUBERT et M. LAFOURCADE. 2009)



Extraits d'onto-terminologies marqués par leurs domaines.



Extraits d'onto-terminologies avec quelques liens typés (médecine). On remarquera la polysémie du terme "température", terme associé à plusieurs voisins selon les domaines. Les liens sont parfois difficilement étiquetables.

2.2 Onto-terminologies: création

La création et/ou l'enrichissement d'onto-terminologies sont des études particulièrement intéressantes pour les langues à faible diffusion, car justement, elles ne possèdent pas de terminologie bien établie, et en outre elles ne sont pas terminologiquement complètes (c'est-à-dire qu'il leur manque des termes pour désigner des pratiques relevant d'un domaine). Mais avant de considérer le problème particulier des langues à faible diffusion nous allons nous intéresser à la création d'onto-terminologies de spécialité.

Une langue de spécialité, utilisée dans un domaine particulier, ou dans une pratique, se comporte souvent, du point de vue lexical, comme une langue à faible diffusion.

1. Elle est partagée par une communauté restreinte (celle qui relève du domaine)
2. Elle est obligée d'emprunter la majorité de son vocabulaire à une langue de plus grande diffusion (la langue générale)
3. Elle a un vocabulaire propre
4. Elle a des acceptions propres des termes du vocabulaire général (voir les figures plus haut
5. Elle ne lésine pas sur les emprunts aux autres langues.

Dans les langues à bonne diffusion, les chercheurs se sont penchés sur la création d'onto-terminologies de spécialité, plus particulièrement dans le domaine médical, celui des sciences agro-alimentaires, celui de l'économie, du droit, etc. Des thésaurii tels que MESH (V.J-L.SCHULMAN, 1997) et ses ontologies dérivées sont actuellement des standards dans le domaine des onto-terminologies médicales, pour la communauté mondiale.

Trois façons de procéder existent pour créer ces onto-terminologies, mais elles partent toujours d'une même amorce, celle de l'existence d'un schéma conceptuel, même restreint, permettant de "commencer" l'arborescence.

1. La première méthode consiste à créer l'arborescence "à la main", à partir d'interactions avec des experts du domaine. Si ce type de fonctionnement a prévalu jusqu'à la fin des années quatre-vingt dix, depuis l'avènement de la très forte disponibilité des corpus numérisés, elle semble être abandonnée, en raison de son coût prohibitif.
2. La deuxième méthode consiste à fouiller des textes de la spécialité et à créer des liens entre termes, puis à décider de leur conceptualisation (V.V.CEAUSU et S.DESPRE ,2005), (V.M.ROCHE ,2003). Cette méthode est incrémentale, et nécessite l'intervention d'un expert dans la boucle pour valider les "trouvailles" des méthodes automatiques. La même technique est utilisable pour peupler l'onto-terminologie (augmentation des termes par des relations de synonymie) (V. J.MAKKI et al. 2008).
3. La troisième consiste à produire une onto-terminologie de manière contributive par don (crowdsourcing). Les dépôts sont faits puis modérés (à la manière de Wikipedia). C'est le cas des différentes déclinaisons de Jeux de Mots (V.M.LAFOURCADE et A.JOUBERT, 2009). Cette technique a l'énorme avantage de puiser dans le savoir collectif, mais l'inconvénient de nécessiter une trop forte modération, pour minimiser le bruit (introduction d'éléments erronés).

2.3 Les onto-terminologies de spécialité et les langues à faible diffusion

Si les langues de spécialité se comportent comme des langues à faibles diffusion, ces dernières ont ceci de particulier qu'elles ne comportent que fort peu de langues de spécialité en tant que telles : elles sont obligées de les emprunter à une langue de plus forte diffusion.
Ainsi, le langage médical, juridique, scientifique ou économique est plus ou moins implanté dans une langue à faible diffusion, d'autant plus qu'elle a la concurrence, à l'écrit, d'une langue de plus forte diffusion dans le même pays.

Une langue comme l'allemand, par exemple, en dépit de son enracinement fort en Europe, est obligée de très fortement cohabiter avec l'anglais pour les domaines de spécialité. C'est encore plus vrai pour le Danois, le Suédois, et autres langues scandinaves.

Le problème donc qui apparaît, est que, pour une langue moyennement ou faiblement diffusée, les ressources termino-ontologiques présentent des "manques" certains de termes, soit techniques et/ou technologiques, soit dans des domaines de spécialité (allant des savoirs académiques fondamentaux comme la philosophie ou les mathématiques aux savoirs des métiers).

L'idée est alors d'utiliser les mêmes techniques que pour la création ou l'augmentation de dictionnaires bilingues, ou le peuplement d'ontologies, pour tenter de pallier ces manques.

On va se situer dans un cas idéal, celui où l'on possède les éléments précédemment répertoriés, à savoir :

1. Des dictionnaires bilingues entre la langue L (objet d'étude) et une langue L', mieux dotée,
2. Des corpus dans la langue L'
3. Des corpus dans la langue L
4. Des dictionnaires monolingues en L'

Les corpus (2) et (3) ne sont pas forcément alignés.
Les dictionnaires bilingues (1) ne sont pas forcément complets.

On rajoutera éventuellement des fragments d'onto-terminologie en *L'* (nommés *F'*) pour le domaine que l'on veut traiter. Si ces fragments n'existent pas, il faut les produire selon les techniques indiquées en 2.2.

L'algorithme de création de l'onto-terminologie *F* en langue *L* est le suivant :

Pour toute étiquette de sommet de F' (en L')
traduire en L à l'aide de (1)
si le terme en L n'existe pas alors conserver le terme de L'
Pour toute étiquette de relation de F '(en L)
copier cette étiquette telle quelle

Une fois cette création terminée, on peut se retrouver face à au moins deux problèmes :

1. beaucoup trop de termes de L' conservés (mots étrangers)
2. des sommets différents de F possèdent la même étiquette. Cela arrive souvent : par exemple, c'est le même terme qui, en arabe, désigne "température", "chaleur" et "fièvre", alors que ces deux derniers "concepts" relèvent de branches conceptuelles différentes dans notre fragment en français (voir figure) (idem en anglais : «température", "fever" et "heath").

Dans le deuxième cas, c'est très simple, il suffit de fusionner les sommets. Ainsi le terme "température" fusionnera dans l'arborescence avec "chaleur" et "fièvre", et l'onto-terminologie arabe sera moins profonde dans ce cas que les onto-terminologies française et anglaise.

Dans le premier cas, on peut y remédier à l'aide des autres éléments répertoriés, sous la forme suivante :

Pour tout terme t' de L' conservé dans F
Rechercher sa présence dans les corpus de L' (ressource (2)) :
S'il est présent,
Étudier ses co-occurences (mesures d'information mutuelle, coefficient de Dice, etc...)
Constituer un graphe de ses voisins dans ces corpus
Comparer avec le graphe de ses voisins dans F'

Si les graphes ne s'apparient que très peu alors
« t' » est probablement polysémique
Rechercher dans le dictionnaire monolingue de L'(ressource (4)) le texte d'explication le plus proche de l'encadrement de t' dans F'
Proposer (avec des experts) la paraphrase en L de la meilleure explication de t' comme étiquette de remplacement
Si les graphes s'apparient bien alors
t' est plutôt monosémique
Proposer une traduction "libre" (avec experts) de t' dans L et l'affecter comme étiquette de remplacement ou construire un néologisme à partir des voisins traduits de t', en utilisant les règles de morphologie dérivationnelle de L
Si t' absent des corpus de L'
Alors pour tout terme traduit, voisin de t' dans F
Rechercher dans les corpus de L (ressource (3)) ce terme
S'il est présent,
Étudier ses co-occurrences (mêmes méthodes)
Proposer une augmentation de l'onto-terminologie F par adjonction de ses possibles voisins
Étudier une révision possible de F (suppression du nœud étiqueté t' et remplacement de son voisinage)

On voit bien que cet algorithme est interruptible en plusieurs endroits en fonction des décisions humaines et que la présence d'experts est indispensable. Néanmoins, il a l'avantage de fournir des pistes et de détecter les manques.

On peut améliorer le procédé en :

1. augmentant le nombre de corpus en langue L, et en langue L'
2. augmentant le nombre de paires de langues avec L, si cette dernière cohabite avec d'autres langues (c'est le cas par exemple de plusieurs langues de l'Inde, (V. M.G.A.MALIK et al. 2008) avec des corpus alignés, et des techniques d'appariement
3. utilisant le "crowdsourcing" (mot emprunté) comme mode de validation, en particulier dans des domaines technologiques. On peut par exemple faire appel au public pour choisir une traduction euphonique de termes comme "téléphonie mobile" et demander, à la manière de Jeux de Mots, à quelles "idées" il les associerait pour créer des onto-terminologies liées à la technologie de la vie moderne.

Des termes peuvent résister à toute transformation dans la langue d'étude : il s'agit souvent de noms de marques (e.g. Smartphone, Ipad, Google, Facebook...), parfois d'acronymes techniques (comment traduire IRM /MRI en arabe ou tamazigh par exemple, on finit par se lasser de la décomposition systématique de l'acronyme...).

D'autres termes peuvent être préférés dans leur langue d'origine (souvent l'anglais) à d'éventuels termes en langue L, a fortiori lorsque L ne fournit que des périphrases pour désigner le concept. La validation de la terminologie est alors dépendante de la popularité, et pas seulement de l'intervention experte. C'est pourquoi, il peut sembler nécessaire de mélanger à la fois les interventions expertes et la légitimité populaire. Cette dernière peut être interpellée de deux façons :

- par le biais du nombre de touches ("hit") des moteurs de recherche, ou de la fréquence sur Internet, cette fois-ci dans la langue L (objet d'étude), si toutefois cette dernière possède au moins une diffusion minimale.
- par le biais de la présentation à des utilisateurs, soit à travers des réseaux sociaux (le J'aime de FaceBook, de Google +...), des fora de discussion, soit en interpellation directe.

Conclusion

Ce premier tour d'horizon a essentiellement pour but de montrer que l'on peut améliorer la diffusion de certaines langues en améliorant leurs ressources numérisées. Les ressources les plus importantes sont bien entendu les dictionnaires, monolingues ou multilingues, et les documents patrimoniaux. Nous en avons rapidement parlé dans la première partie de cette communication.

Si l'on souhaite améliorer cette fois-ci l'exploitation de cette mémoire et de possibles productions textuelles qui pourraient circuler sur internet, il faut être en mesure d'indexer, de rechercher de l'information, d'apparier des bouts de textes, de défricher des grands réservoirs textuels. Pour cela, les onto-terminologies ont leur rôle à jouer. Elles sont aussi importantes pour pointer du doigt et, si possible, combler les lacunes lexicales dans les métiers, les techniques, les spécialités.

Plusieurs techniques sont possibles, et nous avons rapidement décrit celles qui étaient au plus près de l'unité "lexicale", utilisant deux types de ressources : des dictionnaires (numérisés) et des corpus. La population et l'augmentation des onto-terminologies sont des pratiques qui commencent aujourd'hui à être éprouvées, et les procédures que nous avons proposées (dans la deuxième partie) sont une mise en relation de différentes techniques déjà connues. C'est la mise en commun, dédiée à une tâche de promotion d'une langue faiblement dotée en ressources onto-terminologiques, qui pourrait avoir un caractère plus "original" que les techniques elles-mêmes, lorsqu'elles sont examinées séparément.

De plus, nous avons restreint notre discours aux seules ressources lexicales, conditions hélas nécessaires mais guère suffisantes. L'idéal, pour augmenter la participation de l'automatisation, serait d'avoir à sa disposition, pour la langue L' (souvent co-occurrente avec
L, objet d'étude) des outils de TAL qui traiteraient des aspects structurels, à savoir des étiqueteurs morphosyntaxiques, et des analyseurs. Ces outils permettraient de mieux rendre compte de termes de longueur supérieure à ce que peut faire une technique statistique relevant du modèle des n-grammes dont les résultats les plus précis se situent autour d'une granularité aux alentours de 2 ou 3.

Il ne faut cependant pas se voiler la face : la complexité de toute langue humaine est telle que soit on considère le discours comme un "sac de mots" (techniques quantitatives), on ne s'intéresse pas aux structurations et aux mises en perspective, et là, la machine donne des résultats susceptibles d'impressionner, soit on cherche à rendre compte des constructions, et on se heurte aux difficultés de traduction, sous forme de programmes, des processus inhérents à la complexité grammaticale.

Il est donc de bon ton de restreindre ses ambitions et de considérer que ce qui sera gagné, en peuplant ou comblant quelques lacunes dans des onto-terminologies, est toujours fort utile, puisqu'il aura pour objectif d'améliorer l'exploitation des réservoirs de corpus numérisés, ainsi que la reconnaissance et la diffusion de textes, et par là même, de langues dont la conservation eût été autrement plus aléatoire.

Bibliographie

1) BACHIMONT.B. 2013 *: Patrimoine numérique : technique et politique de la mémoire,* Bry-sur- Marne, INA, p.280

2) CEAUSU V., DESPRÈS S, 2005 : « Fouille de textes pour orienter la construction d'une ressource terminologique. » *Proceedings of EGC 2005:* 239-244.

3) DETEY S., DURAND J., LAKS B., LYCHE C., NOUVEAU D., 2010: «Voix de la francophonie, éducation langagière et corpus numérisé : PFC-EF, des ressources pour la didactique du français » i*n DETEY S., DURAND J., LAKS B. & Ch. LYCHE (eds), Les variétés du français parlé dans l'espace francophone.* Ophrys, Collection L'Essentiel Français.

4) FERRARI S., PRINCE V., 1996 : « Création et extension automatiques de dictionnaires terminologiques multilingues spécialisés à partir de corpus monolingues ». *Proceedings of the International Conference on Natural Language Processing and its Industrial Applications (NLPIA).* Moncton (Canada). Vol 1. Pp 79-85.
5) HARRIS Z., 1954: « Distributional Structure ». *Word* 10:2/3. Pp 146-162.
6) IHADJADENE M. (dir.), ZACKLAD M. (dir.), ZREIK K. (dir.) ,2011: *Document numérique entre permanence et mutation : actes du 13e Colloque international sur le document électronique (*CiDE 13), 16- 17 décembre 2010, INHA, Paris, Europia, 250 p.
7) JOUBERT A., LAFOURCADE M, 2008 : « JeuxDeMots : un prototype ludique pour l'émergence de relations entre termes. » *Proceedings of JADT'2008,* Lyon, France.
8) LAFOURCADE M., JOUBERT A, 2009 : « Sens et usages d'un terme dans un réseau lexical évolutif. » *Proceedings of TALN'09,* Senlis, France.
9) MAKKI J., ALQUIER A-M., PRINCE V., 2008: « An NLP-based ontology population for a risk management generic structure», *Proceedings of CSTST 2008:* 350-355.
10) MALIK, M. G. A., BOITET, C, 2008. BHATTCHARRYA, P. «Hindi Urdu Machine Transliteration using Finite-state Transducers», *The 22nd. International Conference on Computational Linguistics (COLING 2008),* Manchester, UK. pp. 537-544.
11) MILLER G. A., 1995: « WordNet: A Lexical Database for English. » *Communications of the ACM* Vol. 38, No. 11: 39-41.
12) PLOUX, S., VICTORRI B., 1998 : « Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes », *TAL,* Vol 39/1, pp. 161-182.
13) PLOUX, S., BOUSSIDAN, A., JI, H. «The Semantic Atlas: an Interactive Model of Lexical Representation. » *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10),* Valletta, Malta. 2010
14) ROCHE, M., 2003 :« Extraction paramétrée de la terminologie du domaine. » *Proceedings of EGC 2003:* 295-306.
15) ROSMORDUC, S. « Le codage informatique des langues anciennes, le cas des hiéroglyphes égyptiens », *Documents Numériques* 6 :3-4, p. 211–225. 2002
16) SALTON G., WONG A., YANG C.S, 1975: «A vector space model for automatic indexing» *Communications of the ACM,* 18, 11, pp. 613-620.
17) SCHULMAN, J-L. , 1997: «MeSH on the Web». *NLM Tech Bull.;* September-October; 298.
18) SEGURA J., PRINCE V., 2011: «Alignment Memories: A Useful Tool to Handle Phrase Alignment Bottleneck », *Proceedings of CLA'2011: Computational Linguistics-Applications Conference.*
19) SERASSET G., BOITET C., 1999: « UNL-French Deconversion as Transfer and Generation from an Interlingua with Possible Quality Enhancement through Offline Human Interaction. » *Proceedings of the Machine Translation Summit VII.*
20) TANAKA K., PRINCE V.1995 «Dictionnaires bilingues incrémentaux utilisant des corpus monolingues.(Incremental Bilingual Dictionaries using Monolingual Corpora). » Proceedings of the IV Journées Scientifiques *" Lexicomatique et Dictionnairiques".* AUPELF-UREF (ed.) Lyon.