Une nouvelle architecture intégrant les données lexicales générales, terminologiques et " situées ": Pivax-3

Ying Zhang, Valérie Bellynck, Mathieu Mangeot, Christian Boitet

▶ **To cite this version:**

Ying Zhang, Valérie Bellynck, Mathieu Mangeot, Christian Boitet. Une nouvelle architecture intégrant les données lexicales générales, terminologiques et " situées ": Pivax-3. Terminology & Ontology: Theories and applications TOTh 2017, Jun 2017, Chambéry, France. hal-01712276

HAL Id: hal-01712276
https://hal.science/hal-01712276

Submitted on 24 Jan 2019

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.

Une nouvelle architecture intégrant les données lexicales générales, terminologiques et « situées » : Pivax-3

Ying Zhang[1]*, Valérie Bellynck[2]*,
Mathieu Mangeot[3]*, Christian Boitet[1]*

* GETALP - LIG, Bâtiment IMAG
700 avenue Centrale, Domaine Universitaire, 38401 Saint Martin d'Hères
[1] Université Grenoble Alpes
[2] Grenoble-INP
[3] Université Savoie Mont Blanc
ying.zhang@imag.fr
valerie.bellynck@imag.fr
mathieu.mangeot@imag.fr
christian.boitet@imag.fr

Abstract. Le point de départ de la nouvelle architecture proposée ici a été un besoin précis d'un partenaire industriel de l'entreprise Lingua et Machina, à savoir la gestion d'un certain type d'abréviations, les acronymes, pour certains clients.
Les acronymes en question sont des unités lexicales souvent terminologiques, mais pas toujours. Un même « prolexème » (collection d'unités lexicales synonymes dans une certaine situation, spatio-temporelle et/ou technique et/ou entrepreneuriale) n'a pas nécessairement de réalisation de type acronyme (ou abréviation, ou apocope, ou mot-valise) dans différentes langues. Du point de vue informatique, il faut pouvoir traiter ces unités dans la même base de données que les termes techniques et que les termes généraux. Nous avons proposé un modèle pour réaliser l'intégration de ces 3 types d'unités lexicales dans une même base de données lexicales. Nous avons pu l'implémenter complètement en Pivax-3/Jibiki-2, et produire un démonstrateur.

1. Introduction

Ce travail se situe principalement dans le domaine de la lexicographie computationnelle, qui a pour objectif d'outiller la mémorisation informatique des informations lexicales et leur accès par des utilisateurs et par des applications. Il s'agit donc de l'informatisation de l'accès au sens des termes à travers leurs formes écrites.

Une difficulté apparaît pour les sens pouvant avoir plusieurs formes, évoluant potentiellement dans le temps, comme les abréviations (sigles, apocopes, mots-valise, acronymes…). Nous nous intéressons ici aux acronymes. Le besoin a été identifié dans le cadre d'une thèse CIFRE [Zhang, 2016] avec Lingua et Machina [Lingua&Machina, 2015] (L&M), pour un partenaire industriel de l'entreprise.

Les acronymes en question sont des unités lexicales souvent terminologiques, mais pas toujours. Un même «prolexème» (collection d'unités lexicales synonymes dans une certaine situation, spatio-temporelle et/ou technique et/ou entrepreneuriale) [Tran, 2006] n'a pas nécessairement de réalisation de type acronyme (ou abréviation, ou apocope, ou mot-valise) dans différentes langues. De plus, du point de vue informatique, il faut pouvoir traiter ces unités dans la même base de données que les termes techniques et que les termes généraux.

Notre but est non seulement d'intégrer les acronymes dans la gestion de terminologies multilingues, mais aussi de traiter l'association de plusieurs termes d'une même langue à un même référent : *Jean-Paul II* et *Karol Jozef Wojtyla* en français, ou en anglais *John Paul II* et *Karol Jozef Wojtyla*. De même, certains liens évoluent avec le temps : le pape désignait *Jean-Paul II* en 2004 et *Benoît XVI* en 2012. Pour la gestion des acronymes, un terme et son acronyme peuvent désigner le même référent. Par exemple, pour la ville de Saint-Martin-d'Hères, on peut trouver : *Saint Martin d'Hères*, *St Martin d'Hères*, *Saint-Martin-d'Hères*, *SMDH*, ou *SMH*.

Dans un contexte multilingue d'aide aux traducteurs, réviseurs et post-éditeurs, ~~il s'agit~~ de proposer, pour un terme donné dans une langue donnée, l'équivalent le plus précis possible dans chaque langue cible visée.

Pour trouver une solution générique, qui permette à une base lexicale de contenir tous les types d'unités lexicales, nous avons défini un cadre théorique composé d'une nouvelle macrostructure[1] basée sur des concepts existants et

1 Définition donnée au 3 ci-dessous.

sur la définition de nouveaux concepts. Nous avons pu ~~installer~~ le premier prototype à l'aide d'un outil générique de gestion de bases lexicales (JIBIKI-2).

2. Conception globale

Les idées présentées dans cet article sont basées sur la théorie de la cognition située, introduite dans les travaux de Coseriu [Coseriu, 1998 ; Coseriu, 2001]. Citons ici la définition trouvée dans WIKIPEDIA :

> Situated cognition is a theory that posits that knowing is inseparable from doing [Brown *et al.*, 1989] by arguing that all knowledge is situated in activity bound to social, cultural and physical contexts [Greeno & Moore, 1993].

La difficulté de l'unification dans une même base lexicale de lexèmes (simples ou composés) généraux, de termes techniques, et de prolexèmes (abréviations, noms propres et entités nommées) semble provenir du fait qu'ils sont plus ou moins « situés ».

Un lexème général comme *chemin*, ou *boire*, ou *vite* a une acception (sens en usage) qui ne dépend pas d'un contexte d'interprétation particulier. Nous dirons qu'il est **général**, ou non situé.

Un terme technique comme *ampoule* a des sens qui dépendent de domaines particuliers (électricité, médecine, religion), mais pas du lieu ni du temps. Nous dirons qu'il est **terminologique**.

Une abréviation comme *CNAM* peut signifier « Caisse Nationale d'Assurance Maladie » ou « Conservatoire National des Arts et Métiers ». Un toponyme comme *Rome* renvoie à une ville, certes, mais s'agit-il de la capitale de l'Italie, de Rome (New-York), ou d'autres villes de ce nom aux USA ? Pour l'interpréter correctement, il faut disposer de la situation, ou au moins de paramètres pertinents de la situation. Nous dirons que *CNAM* et *Rome* sont **situés**.

Enfin, le sens des noms de société, ou des noms de personne, comme *George W. Bush*, dépend souvent d'une description encore plus précise du contexte. Même si l'on ajoute *président des USA*, il y en a eu deux de ce nom… et cette qualité leur reste attachée. En fait, le sens de beaucoup d'entités nommées dépend non seulement du domaine et du lieu, mais aussi du temps. On peut dire que les noms de personne et ce qu'on appelle depuis quelques années les entités nommées sont **très situés**. Par exemple, *la capitale de la RFA* a été Bonn, puis Berlin à partir de 1991.

Un prolexème est, par définition, une classe d'équivalence d'expressions snonymes par rapport à une certaine situation. Ainsi, P1 = *{président des USA, Georges W. Bush#2}* est valide par rapport à 2005, et P2 = *{président des USA, Barack Obama#1}* est valide par rapport à 2015.

Nous avons défini le terme «**situement**»[2] comme la qualité d'être situé, et parlons dès lors du «**degré de situement**» d'un lexème.

Chaque type de vocable (entrée d'un dictionnaire usuel [Polguère, 2002]) correspond à un degré de situement : général, terminologique, situé, ou très situé.

Pour bien expliquer les degrés de situement, nous donnons ici quelques exemples.

Tab. 1 - Exemples des différents degrés de situement

Mot-vedette	Degré de situement	Descripteur de situement	Signification
sémaphore	général	sémaphore#général	signalisation maritime
	situé	sémaphore#situé?lieu=Béar	sémaphore du cap Béar (Chemin du Cap Béar, 66660 Port-Vendres)
	termino-logique	sémaphore#terminologique?domaine=informatique	en informatique, dispositif de verrouillage de ressources
diabolo	général	diabolo#général	instrument de jonglage
	termino-logique	diabolo#terminologique?domaine=médecine	aérateur en ORL
président	général	président#général	chef d'État, PDG d'une compagnie, président d'un tribunal, etc.
	très situé	président#situé?lieu=États-Unis&année=2015	Barack Obama
	très situé	président#situé?lieu=États-Unis&année=2005	George W. Bush

2 Ce terme «situement» semble naturel, car il est solution de plusieurs équations analogiques [Lepage, 2000] comme : `situer:situement::figer:figement::dénué:dénuement`. On parle de même du «degré de figement» d'une collocation.

Une « base lexicale métier » est une base lexicale adaptée à ~~différents~~ métier~~s~~ (traducteur, lexicographe, terminologue, enseignant…), et à différents logiciels, pour certains domaines et certaines tâches. Une base lexicale métier unifie tous les types d'unités lexicales, et en particulier tous les degrés de « situement ».

3. Implémentation

La macrostructure d'une base de données lexicale est la description de son architecture générale, c'est-à-dire des types de ses volumes et de leurs relations. PIVAX [Nguyen & Boitet, 2009] est une macrostructure à trois niveaux : les volumes de lexies (une lexie est un sens de mot), les volumes d'axèmes (un axème est une acception monolingue) et un unique volume d'axies (une axie est une acception interlingue). Cette macrostructure a été définie et implémentée par H.-T. Nguyen dans le cadre de sa thèse [Nguyen, 2009]. Elle permet d'implémenter plusieurs volumes monolingues pour chaque langue naturelle ou artificielle.

Pour traiter ~~des~~ bases lexicales métier, nous avons conçu une nouvelle macrostructure, intégrée à PIVAX-3. Elle contient deux couches, une couche « basique » et une couche « pro ». Dans la couche basique, nous gérons trois types de volume comme PIVAX. Dans la couche « pro », nous gérons deux nouveaux types de volume : les volumes de prolexèmes (un prolexème regroupe les lexies monolingues qui représentent le même sens mais dont la réalisation syntaxique est différente) et le volume des proaxies (une proaxie regroupe les prolexèmes de langues différentes partageant un même sens).

Grâce à la couche basique, nous pouvons relier les lexies qui se correspondent exactement, comme l'acronyme français *ONU*, relié à l'acronyme anglais *UN*.

Grâce à la couche « pro », nous pouvons proposer les traductions de lexies de même sens dans les langues cibles, avec jusqu'à 3 degrés de précision. Par exemple, en chinois, il y a un seul mot 联合国 (lián hé guó) pour ce sens, et il n'existe pas d'acronyme. Donc on peut toujours proposer le même terme 联合国 comme traduction de *UN* et comme traduction de *United Nations*. Voir la Fig. 1.

L'utilisation d'étiquettes sur les liens (on parle de « liens riches ») a pour but de proposer les traductions les plus précises possibles. Par exemple, en japonais, 国際連合 (kokusai-rengō) est la lexie qui a exactement le même sens que *United Nations*, et son acronyme est 国連 (kokuren). Cet acronyme utilise

le premier et le troisième kanji de ce mot (composé), ce qui est différent des initiales de la lexie de définition (le cas de *ONU* et de *UN*).

Il arrive que, dans une langue, il y ait deux acronymes de même sens pour une expression polylexicale, l'un correspondant à l'acronyme des initiales, l'autre correspondant à une sélection de caractères ou de mots. Donc, nous avons décidé de ne pas relier ces deux acronymes de types différents à une même axie, mais seulement à une même «proaxie». Par contre, comme ce sont des acronymes, pour traduire *ONU*, 国連 est meilleur que 国際連合.

Comme ~~notre étiquette est une~~ étiquette libre, pour chaque base de données lexicales de type PIVAX-3, il faut bien définir les étiquettes utilisées pour conserver la cohérence.

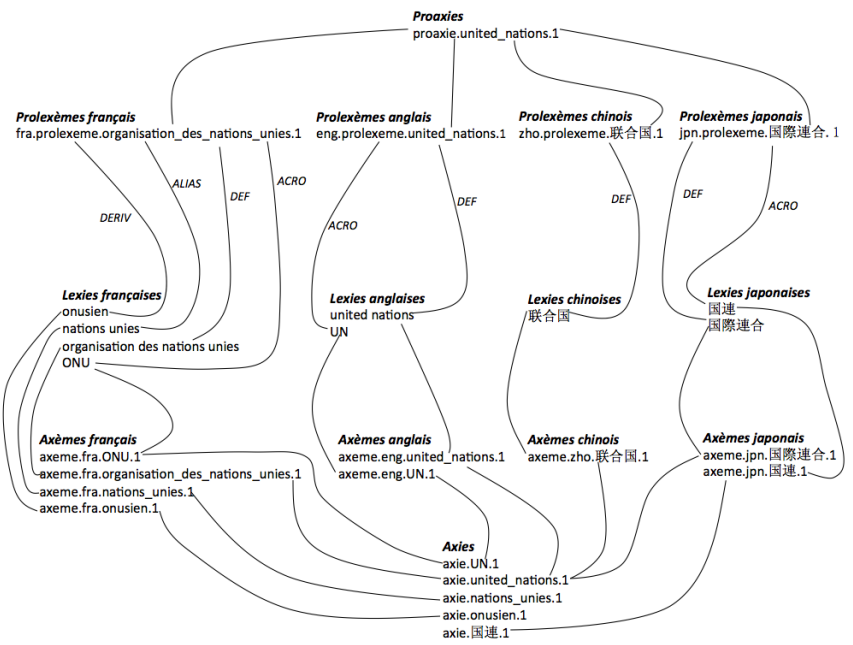
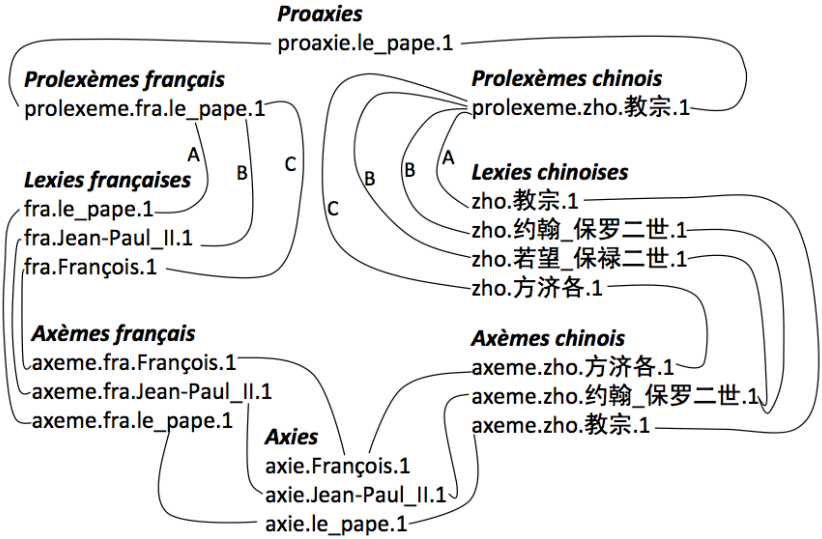


Fig. 1 - Exemple de liens dans PIVAX-3



Descripteur de situement (étiquettes)

A	le_pape#général
B	le_pape#situé ? date=16101978- 02042005
C	le_pape#situé ? date=19042005- 28022013

Fig. 2 – Modélisation de l'exemple « le pape » dans PIVAX-3

Pour traiter les lexèmes « très situés », nous prévoyons de prendre en compte également les quatre dimensions du diasystème basé essentiellement sur celles qu'Eugenio Coseriu a proposées : diachronique (variété dans le temps), diaphasique (variété concernant les finalités de l'emploi), diatopique (variété dans l'espace), et diastratique (variété relative à la stratification socio-culturelle).

Pour ces cas complexes, nous utilisons le descripteur de situement comme étiquette. C'est le cas dans la modélisation de l'exemple « *le pape* » dans PIVAX-3, donnée dans la FIG. *2*.

Les deux lexies chinoises *zho.(约翰_保罗二世.1)* et *zho.(若望_保禄二世.1)* sont synonymes. On peut les échanger dans tous les contextes. On propose donc ces deux traductions, au premier niveau de traduction, pour *fra.Jean-Paul_II.1*.

4. Trois niveaux de traduction

Nous proposons trois niveaux de traduction, classés selon la précision obtenue.

1. Le système trouve une lexie directement, en passant par le volume des axèmes et par le volume des axies. C'est le premier niveau de traduction, et le plus précis. Par exemple, la traduction de *UN* en français est *ONU*.
2. Le système cherche le lien dans le volume des prolexèmes de la langue source avec une étiquette. Puis il parcourt le volume des proaxies, et ensuite le volume des prolexèmes et les volumes des lexies des langues cible. Il trouve une lexie avec la même étiquette. C'est le deuxième niveau, dit niveau intermédiaire. Par exemple, en japonais, 国連 est meilleur que 国際連合 pour traduire *UN*, parce que ces mots portent l'étiquette `acronyme`.
3. Le système trouve les lexies en passant par les prolexèmes et les proaxies, sans prendre en compte l'étiquette. Ces lexies proposées constituent le troisième niveau, le moins précis. Par exemple, on trouvera la traduction en chinois 联合国 pour *UN*.

La quantité de lexies contenues dans le résultat augmente suivant les niveaux de traduction, du premier vers le troisième. C'est-à-dire qu'on a :

$$\{traductions_1er_niveau\} \subseteq \{traductions_2e_niveau\} \subseteq \{traductions_3e_niveau\}$$

Pour faciliter la lecture dans l'interface Web, nous avons décidé :

1. d'afficher l'étiquette, la langue et le mot-vedette dans le 1er et le 2e niveau,
2. d'afficher tous les détails (phrases exemples, définitions, POS, etc.) dans le 3e niveau, y compris les lexies du même prolexème de la langue source, et
3. de ne pas afficher la traduction dans le 2e niveau si elle a déjà été trouvée et est déjà affichée dans le 1er niveau.

5. Outils utilisés

Pour implémenter la macrostructure de Pivax-3, nous avons utilisé la plate-forme Jibiki, qui permet la construction de sites Web contributifs dédiés à la construction et à la gestion de bases lexicales multilingues.

Jibiki a été développée principalement par Mathieu Mangeot et Gilles Sérasset [Mangeot & Sérasset, 2001]. Elle a été utilisée dans divers projets (projet Lexalp [Sérasset, 2008], projet Papillon [Mangeot & Thevenin, 2004], projet GDEF [Mangeot & Chalvin, 2006], etc.). Le code est disponible en source ouvert et téléchargeable gratuitement par SVN sur ligforge.imag.fr. Jibiki permet l'import, l'export, l'édition, la modification et la recherche dans des bases lexicales. On peut aussi gérer les contributions.

La version 1 de Jibiki présentait deux limitations : on ne pouvait pas décrire des liens entre des volumes différents, ni ajouter des attributs (poids, étiquette, volume cible, etc.) sur les liens.

Pour traiter des liens plus complexes, nous avons enrichi la plate-forme Jibiki, obtenant Jibiki-2 [Zhang *et al*., 2014]. Avec cette nouvelle version, plusieurs informations peuvent être portées par chaque lien :

- l'identifiant de l'entrée source.
- l'identifiant de l'entrée cible.
- l'identifiant de l'élément Xml de l'entrée source contenant le lien. Par exemple, le numéro de sens dans le cas d'une entrée polysémique avec un lien de traduction pour chaque sens. Cela permet de retrouver précisément l'origine du lien.
- le nom du lien. Celui-ci est utilisé pour distinguer des liens de types différents dans une même entrée, par exemple un lien de traduction et un lien de synonymie.
- la langue cible (code à trois lettres ISO-639-2/T).
- le volume cible.
- le type de lien. Certains types sont prédéfinis car ils sont utilisés par les algorithmes de calcul des liens riches (traduction, axème, axie), mais il est possible d'en utiliser d'autres.
- une étiquette dont le texte est libre.
- un poids dont la valeur doit être un réel dans]-10,+10[.

6. Démonstrations

Nous terminons en donnant trois démonstrations de consultation.

6.1. Consultation du terme UN de l'anglais vers les autres langues

Il s'agit d'une consultation pour les trois niveaux de traduction.

Tab. 2 - Trois niveaux de traduction : terme UN de l'anglais vers toutes les langues

Niveau	**Lexies trouvables en théorie**			
	Français	Anglais	Chinois	Japonais
1	*ONU*			
2	*ONU*			国連
3	*ONU, onusien, Nations unies, Organisation des nations unies*	*Q='UN'*	联合国	国連, 国際連合

Niveau	**Lexies trouvées et affichées par l'interface**			
	Français	Anglais	Chinois	Japonais
1	*ONU*			
2				国連
3	*ONU, onusien, Nations unies, Organisation des nations unies*	*UN, United Nations*	联合国	国連, 国際連合

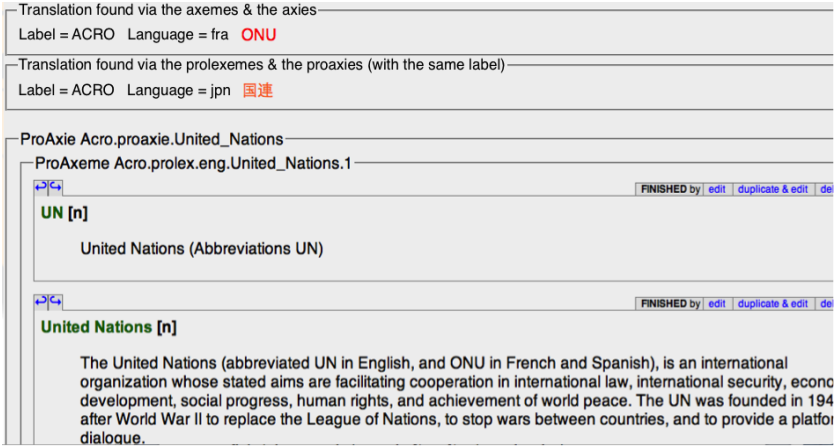


Fig. 3 - Terme UN de l'anglais vers toutes les langues

6.2. Consultation du terme 国連 du japonais vers les autres langues

Il s'agit d'une consultation aux 2e et 3e niveaux de traduction.

Tab. 3 - Trois niveaux de traduction : terme 国連 du japonais vers toutes les langues

Niveau	**Lexies trouvables en théorie**			
	Français	Anglais	Chinois	Japonais
1				
2	*ONU*	*UN*		
3	*ONU, onusien, Nations unies, Organisation des nations unies*	*UN, United Nations*	联合国	Q='国連'

Niveau	**Lexies trouvées et affichées par l'interface**			
	Français	Anglais	Chinois	Japonais
1				

Niveau	Lexies trouvées et affichées par l'interface			
	Français	Anglais	Chinois	Japonais
2	*ONU*	*UN*		
3	*ONU, onusien, Nations unies, Organisation des nations unies*	*UN, United Nations*	联合国	国連, 国際連合

Fig. 4 - Terme 国連 *du japonais vers toutes les langues*

6.3. Consultation du terme onusien du français vers les autres langues

Il s'agit d'une consultation au 3e niveau de traduction.

Tab. 4 - Trois niveaux de traduction : terme onusien du français vers toutes les langues

Niveau	**Lexies trouvables en théorie**			
	Français	Anglais	Chinois	Japonais
1				
2				
3	*Q='onusien'*	*UN, United Nations*	联合国	国連, 国際連合

Niveau	**Lexies trouvées et affichées par l'interface**			
	Français	Anglais	Chinois	Japonais
1				
2				
3	*ONU, onusien, Nations unies, Organisation des nations unies*	*UN, United Nations*	联合国	国連, 国際連合

Fig. 5 - Terme onusien *du français vers toutes les langues*

Références

Brown, J. S., Collins, A. and Duguid, P. (1989). Situated Cognition and the Culture of Learning. *Educational Researcher*, Vol. 18, pp. 32-42.

Coseriu, E. (1998). Le double problème des unités dia-s. *Les Cahiers dia. Études sur la diachronie et la variation linguistique*, Vol. 1, pp. 9-16.

Coseriu, E. (2001). L'homme et son langage. Louvain - Paris, Peeters, 492 p.

Greeno, J. G. and Moore, J. L. (1993). Situativity and Symbols : Response to Vera and Simon. *Cognitive Science*, Vol. 17, pp. 49-59.

Lepage, Y. (2000). Languages of analogical strings. Proc. *COLING-2000*, Saarbrücken, pp. 488-494.

Lingua&Machina. L'Écrit multilingue dans l'entreprise. http://www.lingua-et-machina.com

Mangeot, M. and Chalvin, A. (2006). Dictionary building with the Jibiki platform : the GDEF case. Proc. *LREC 2006*, Genoa (Gênes), pp. 1666-1669.

Mangeot, M. and Sérasset, G. (2001). Papillon lexical database project : Monolingual dictionaries and interlingual links. Proc. *NLPRS-2001*, Tokyo, pp. 119-125.

Mangeot, M. and Thevenin, D. (2004). Online generic editing of heterogeneous dictionary entries in the Papillon project. Proc. *COLING 2004*, Geneva, 7 p.

Nguyen, H.-T. (2009). Des systèmes de TA homogènes aux systèmes de TAO hétérogènes. *Thèse de doctorat en informatique*, Université Joseph-Fourier - Grenoble I, 236 p.

Nguyen, H. T. and Boitet, C. (2009). Lexical synergy between MT & Translator Aids : PIVAX, a generic online contributive lexical database platform. Proc. *International Conference "Machine Translation 25 Years On"*, Cranfield, 8 p.

Polguère, A. (2002). Notions de base en lexicologie, OLST-Département de linguistique et de traduction, Université de Montréal, 140 p.

Sérasset, G. (2008). La plate-forme » Jibiki » dans le projet LexALP. Proc. *Normazione, armonizzazione e pianificazione linguistica ; Normierung, Harmonisierung und Sprachplanung ; Normalisation, harmonisation et planification linguistique*, Bolzano, pp. 31-47.

Tran, M. (2006). Prolexbase : un dictionnaire relationnel multilingue de noms propres : conception, implémentation et gestion en ligne. *Thèse de doctorat en informatique*, Université de Tours, 171 p.

Zhang, Y. (2016). Modèles et outils pour des bases lexicales « métier » multilingues et contributives de grande taille, utilisables tant en traduction

automatique et automatisée que pour des services dictionnairiques variés. *Thèse de doctorat en informatique*, Université Grenoble Alpes, 202p.

Zhang, Y., Mangeot, M., Bellynck, V. and Boitet, C. (2014). Jibiki-LINKS : a Tool between Traditional Dictionaries and Lexical Networks for Modelling Lexical Resources. Proc. *Workshop on Cognitive Aspects of the Lexicon (CogALex), COLING 2014*, Dublin, pp. 87-98.