Atilf - UMR 7118 CNRS/Université de Lorraine – Nancy

Extraction et édition terminologique
Construction d'une onto-terminologie

Projet **TermITH** ***TERM**inologie et **I**ndexation de **T**extes en sciences **H**umaines* *ANR Contint (Contenus numériques et interactions)*	
Post-doctorat	
Responsables : Laurence Kister – Evelyne Jacquey	Date de début du contrat et durée : Septembre 2014 – 12 mois

Mission

La mission s'articule autour de :

- la structuration et la hiérarchisation des concepts spécifiques aux sciences du langage en une onto-terminologie
- la rédaction de définitions terminologiques pour les termes correspondant aux concepts

Les candidats termes à organiser et à définir sont issus de différentes ressources terminologiques et/ou textuelles :

- référentiel linguistique de l'Inist
- Thesaulangue (thesaurus conçu et maintenu par le centre de documentation de l'Atilf)
- candidats termes extraits automatiquement de textes intégraux (textes du corpus Scientext et sélection de textes de sciences du langage fournis par Open Edition)
- termes extraits de ressources encyclopédiques de type dictionnaires de spécialité et terminologies disponibles
- GTN – gros tas de notions (O. Tremblay, 2009, Thèse) http://olst.ling.umontreal.ca/pdf/OphelieTremblayThese2009.pdf

Les concepts à organiser en fonction des relations qu'ils entretiennent les uns avec les autres feront l'objet d'une définition terminologique. La rédaction de ces définitions s'inspirera du cadre formel du GTN et s'appuiera sur :

- des dictionnaires de spécialité
- des terminologies existantes
- des indices extraits des contextes d'occurrences des termes dans les textes :
  - lexicaux (lexèmes spécifiques issus du calcul de spécificité de Lafon (1980) tel ceux implémentés dans le logiciel TXM ou lexico3)
  - sémantiques issus des définitions lexicographiques des lexèmes contenus dans les contextes d'occurrence des termes
- des référentiels catégorisés sémantiquement conçu par l'Inist (scenario 3 du projet TermITH)

Contexte scientifique

TermITH s'intéresse à l'indexation automatique de textes scientifiques en SHS. Il procède à la détection, la désambiguïsation et la qualification de candidats termes en analysant leurs formes linguistiques dans les textes. Nous nous situons dans une approche allant du texte (réalisations linguistiques de termes dans les textes) aux termes (objet conceptuels) comme, parmi d'autres, (Daille 1994), (Toussaint *et al.* 2001), (Bourigault et Slozdian 1999) et (Bourigault *et al.* 2001). Les travaux menés se situent dans le champ de l'extraction terminologique à partir de textes intégraux appartenant à différents domaines des SHS : la méthodologie est mise au point à partir de textes en sciences du langage, puis testée sur des textes de sciences de l'information et de la communication, d'archéologie, de psychologie et de chimie verte.

Accéder aux réalisations linguistiques des termes dans les textes suppose de les reconnaître comme telles. Parmi les travaux qui abordent cette problématique, une première partie s'appuie sur l'utilisation d'extracteurs automatiques de candidats termes ensuite validés par des experts des différents domaines de spécialités : *Acabit* (Daille 1994 ; 2003), *Yatea* (Aubin et Hamon 2006), *TermoStat* (Drouin 2003) ou encore la plate-forme *TTC-TermSuite* (Daille et *al.*, 2011). D'autres travaux s'intéressent à la validation, à l'extraction et aux relations qui s'établissent entre les termes contenus dans les textes en adoptant une perspective distributionnelle pour identifier les réalisations linguistiques de termes. Les travaux de Daille (2003), Toussaint et *al.* (1998), Namer et Zweigenbaum (2004) ou L'Homme (2004a) utilisent des connaissances relevant de la morphologie dérivationnelle ou constructionnelle. Les travaux de Baneyx et *al.* (2005), Jacques et Aussenac-Gilles (2006), Aussenac-Gilles et Condamines (2009), Kister et Jacquey (2012), Manser (2012), Jacquey et *al.* (2013), Périnet et Hamon (2013) détectent et exploitent des patrons lexico-syntaxiques pour l'identification des relations entre (réalisations linguistiques de) termes. Enfin, les travaux de Grabar et Zweigenbaum (2004), Claveau et L'Homme (2005), Poibeau (2005) ou Condamines et Péry-Woodley (2007) reposent sur l'utilisation de structures sémantiques, textuelles ou discursives.

Notre objectif est l'analyse des contextes d'occurrences des candidats termes afin de sélectionner automatiquement celles qui relèvent d'un usage terminologique et de rejeter les autres. Ainsi, nous procédons à un type particulier de désambiguïsation sémantique que nous appelons *désambiguïsation terminologique*. Comme l'a montré L'Homme (2004b), même si le terme, en tant qu'étiquette de concept dans une terminologie donnée avec une application définie, n'est pas ambigu, ses réalisations linguistiques peuvent l'être. Ceci est particulièrement valable lorsque les *termes* sont des *candidats termes* extraits automatiquement par une plate-forme d'extraction terminologique. Ce phénomène se manifeste aussi lorsque les termes dont on observe les occurrences en texte intégral apparaissent dans des thesauri ou des référentiels terminologiques. Les ambiguïtés peuvent être diverses :

- ambiguïté avec le lexique ou la phraséologie transdisciplinaire : *argument, corpus, définition, énoncé, exemple, objet, référence*

  [+termino] *En chemin, nous avons souligné la grande flexibilité des SN définis pluriel, qui en fait le lieu possible d'une négociation de la **référence** et de la désignation* (Figures et référence plurielle en corpus journalistique - Lecolle M. (2000). Cahiers de grammaire (25))

  [-termino] *Les auteurs font **référence** à [...]* (Validation d'une méthodologie pour l'étude des marqueurs de la segmentation dans un grand corpus de texte - Piérard S. et Begsten Y. (2007). TAL(47/2))

- ambiguïté avec un autre domaine de spécialité : *patient*

  [+termino] *[...] ou plus rarement, à des rôles argumentaux (agent, **patient**, objet,...) [...]* (Les relations sémantiques : du linguistique au formel - Aussenac-Gilles N. et Séguéla P. (2000). Cahiers de grammaire (25))

  [-termino] *es **patients** cérébrolésés [...]* (Nouveaux habits de la lexicographie spécialisée : Intégration de la métaphorique dans le dictionnaire du football - Leroyer P. et Moller B. (2004). EURALEX)

- ambiguïté avec un emploi lexical ou phraséologique de langue générale : *argument, définition, énoncé, expression, objet*

  [+termino] *[...] les **expressions** du type le jour suivant*. (Validation d'une méthodologie pour l'étude des marqueurs de la segmentation dans un grand corpus de texte - Piérard S. et Begsten Y. (2007). TAL(47/2))

  [-termino] ***L'expression** de telle ou telle relation [...]* (Variabilité des outils de TAL et genre textuel : cas des patrons lexico-syntaxiques - Jacques M.-P. et Aussenac-Gilles N. (2006). TAL (47))

Comme le montrent ces quelques exemples de réalisations linguistiques de candidats termes, c'est le contexte au sens large qui nous permet de sélectionner les occurrences relevant d'un emploi terminologique.

Un autre objectif de TermITH est de contribuer à la mise à jour des ressources terminologiques disponibles dans les domaines concernés par le projet. Les sciences du langage constituent le domaine sur lequel la méthodologie de l'ensemble du projet est mise au point et c'est tout naturellement sur ce domaine que porte le sujet du post-doctorat (voir description de la mission). Les ressources terminologiques de références sont Thesaulangue et le vocabulaire de la linguistique de l'Inist. L'optimisation et l'enrichissement des ressources bénéficieront des résultats de deux des tâches du projet TermITH :

- la détection et la désambiguïsation des réalisations linguistiques des termes dans les textes de sciences du langage
- la mise en œuvre de scenarii d'utilisation par l'Inist pour évaluer la qualité des termes extraits des textes en vue de l'alimentation et la mise à jour des ressources terminologiques

Références

Aubin S. et Hamon T. (2006) Improving Term Extraction with Terminological Resources. In *Advances in Natural Language Processin, 5th International Conference on NLP, FinTA*.

Aussenac-Gilles N., Condamines A. (2009). Marqueurs de relations, genre textuel, structures syntaxiques. In Minel, J.-L. (Ed.), *Filtrage sémantique*, 115-149. Paris: Hermes/Lavoisier.

Baneyx A., Malaisé V., Charlet J., Zweigenbaum P. et Bachimont B. (2005). Synergie entre analyse distributionnelle et patrons lexico-syntaxiques pour la construction d'ontologies différentielles. Dans Actes de la conférence *TIA*, 12p. http://estime.spim.jussieu.fr/~jc/Files/BaneyxTIA2005.pdf [page consultée le 3 mars 2014]

Bourigault D., Jacquemin C. et L'Homme M.C. (2001). *Recent Advances in Computationnal Terminology*. John Benjamins :Amsterdam.

Bourigault D.et Slozdian M. (1999). Pour une terminologie textuelle. Revue *Terminologies Nouvelles* 19 (Actes de la conférence *TIA*), 29-32. http://www.rifal.org/cahiers/rint19/rint19.pdf [page consultée le 3 mars 2014]

Claveau V. et L'Homme M.C. (2005). Apprentissage par analogie pour la structuration de terminologies - utilisation comparée de ressources endogènes et exogènes. Dans Actes de la conférence *TIA-2005*, 12 p. http://www.irisa.fr/texmex/people/claveau/publis/Claveau-LHomme-tia05.pdf [page consultée le 3 mars 2014]

Condamines A et Péry-Woodley M.P. (2007). Linguistic markers of semantic and textual relations. In Alamargot, D., Terrier, P. & Cellier, J.-M. (Eds.), *Written documents in the workplace. Studies in Writing*. 3-16. Amsterdam: Elsevier.

Daille B. (1994). Approche mixte pour l'extraction de terminologie : statistique lexicale et filtres linguistiques, Thèse en informatique fondamentale, Université Paris 7.

Daille B. (2003). Conceptual structuring through term variations. In *Proceedings of the ACL2003 Workshop on Multiword Expressions : Analysis, Acquisition and Treatment*, Bond E., Kohonen A. Carthy D.M. et Villalencio A. (eds), 9-16.

Drouin P. (2003). Term extraction using technical corpora as a point of leverage. In *Terminology* 9(1), 99-117.

Grabar N et Zweigenbaum P. (2004). Lexically-based terminology structuring. In *Terminology*, 10(1), 23-54.

Jacques M-P.et Aussenac-Gilles N. (2006). Variabilité des performances des outils de TAL et genre textuel. Cas des patrons lexico-syntaxiques. Dans : *Traitement Automatique des Langues*, 47(1), 11-32. http://www.atala.org/Variabilite-des-performances-des [page consultée le 3 mars 2014]

L'Homme M.C. (2004a). Adjectifs dérivés sémantiques dans la structuration de terminologies. Dans Actes de la conférende *Terminologie, ontologie et représentation des connaissances*, Université Jean-Moulin Lyon-3, 22-23 janvier, 6p. http://olst.ling.umontreal.ca/pdf/lhomme-lyon2003.pdf [page consultée le 3 mars 2014]

L'Homme M.C. (2004b). *La terminologie : principes et techniques*, Montréal : Presses de l'Université de Montréal.

Manser M. (2012). État de l'art sur l'acquisition de relations sémantiques entre termes : contextualisation des relations de synonymie. Dans Actes de la conférence *JEP-RECITA*, 163-175. http://aclweb.org/anthology//F/F12/F12-3013.pdf [page consultée le 3 mars 2014]

Namer F. et Zweigenbaum P. (2004). Acquiring meaning for French medical terminology: contribution of morphosemantics. In Marius Fieschi, Enrico Coiera, and Yu-Chuan Jack Li, editors, *Proceedings 10th World Congress on Medical Informatics*, volume 107 of *Studies in Health Technology and Informatics*, 535-539, Amsterdam, IOS Press. http://www.ncbi.nlm.nih.gov/pubmed/15360870?dopt=Abstract [page consultée le 3 mars 2014]

Périnet A.et Hamon H. (2013). Hybrid acquisition of semantic relations based on context normalization in distributional analysis. Dans Actes de la conférence *TIA*, 113-120. https://lipn.univ-paris13.fr/tia2013/Proceedings/actesTIA2013.pdf [page consultée le 3 mars 2014]

Poibeau T. (2005). Parcours interprétatifs et terminologie. Dans *Actes TIA,* Rouen.

Toussaint Y., Namer F., Daille B., Jacquemin C., Royauté J.et Hathout N. (1998). Une approche linguistique et statistique pour l'analyse de l'information en corpus. Dans *Actes de la conférence TALN*, ATALA, Paris, France.

Profil recherché

- Terminologue
- Linguiste : Sémantique lexicale, Lexicologie

Conditions de recrutement en post-doctorat

- être titulaire d'une thèse depuis moins de 2 ans
- ne pas avoir soutenu sa thèse dans le labo qui propose le post-doctorat

Pour en savoir plus sur l'Atilf

- http://www.atilf.fr

Pour en savoir plus sur TermiTH

- http://www.atilf.fr/ressources/termith/index.php
- http://www.agence-nationale-recherche.fr/suivi-bilan/sciences-et-technologies-de-l-information-et-de-la-communication/contenus-numeriques-et-interactions/fiche-projet-contint/?tx_lwmsuivibilan_pi2[CODE]=ANR-12-CORD-0029

Salaire mensuel 2 500 € brut soit environ **2000 € net**

Lieu de travail
Atilf UMR 7118 CNRS/Université de Lorraine
BP 30687
44, avenue de la libération
54063 Nancy Cedex

Candidature souhaitée pour le 10 mai 2014

Adresser une lettre de motivation accompagnée d'un CV détaillé à Laurence Kister et Evelyne Jacquey
laurence.kister@univ-lorraine.fr et evelyne.jacquey@atilf.fr

Pour toute demande d'information complémentaire : mail ou téléphone
Laurence Kister : 03 54 50 53 43 (lundi et vendredi)
Evelyne Jacquey : 03 54 50 52 94