Une étude terminologique de la communication hypertexte web. Caractéristique du domaine universitaire.

David Reymond, Nathalie Pinède, Véronique Lespinet-Najib, Benoît Le Blanc

Une étude terminologique de la communication hypertexte web.

Caractéristiques du domaine universitaire.

David Reymond

Université du Sud Toulon-Var, I3M EA 3820, Bâtiments C BP 20132 - 83957 La Garde Cedex

david.reymond(at) univ-tln.fr

Nathalie Pinède

Université de Bordeaux, MICA EA 4426, MSHA, Esplanade des Antilles, 33607 Pessac, France

Nathalie.pinede(at) iut.u-bordeaux3.fr

Véronique Lespinet-Najib

Université de Bordeaux, IPB, ENSC, UMR 5218, IMS, 146 rue Léo Saignat 33 076 Bordeaux Cedex

veronique.lespinet(at) ensc.fr

Benoît Le Blanc

Université de Bordeaux, IPB, ENSC, UMR 5218, IMS, 146 rue Léo Saignat 33 076 Bordeaux Cedex

Benoit.leblanc(at) ensc.fr

Résumé

Dans cet article nous décrivons la dynamique temporelle et la représentativité d'un ensemble terminologique. Le vocabulaire est constitué à partir des unités lexicales utilisées pour la construction des menus de navigation des pages d'accueil de sites web d'une zone DNS d'établissements universitaires. Ainsi positionnées, les unités lexicales sont des descripteurs des contenus sous-jacents des sites. Nous montrons que malgré l'évolution très dynamique et tangible des contenus du web, certaines propriétés remarquables émergent quant à la stabilité voire la convergence des choix de ces unités lexicales par les webmasters du domaine.

1 Introduction

Le site web en tant que support de communication hypertextuel relève d'une catégorie de documents numériques particulière (Rouquette, 2009). Les contenus textuels des menus de navigation, constituent le vocabulaire d'une taxonomie de descripteurs décrite par ailleurs (Reymond, 2007 ; Pinède & Reymond, 2010). Après une présentation contextualisée du domaine source de ces unités lexicales hypertexte (ULH), qui soulignera la singularité du corpus en regard de corpus documentaires classiques par les pratiques langagières (Aussenac-Gilles & Condamines, 2004), nous montrerons la représentativité des ensembles terminologiques constitués sur quelques sites seulement, puis leur stabilité en opérant une comparaison diachronique, ces résultats offrant des « indicateurs formels de contextes définitoires et informatifs » (Otman, 1995). Nous appliquons notre démarche aux sites web d'universités dont les sites de nature informative sont considérés comme matures de par leur historicité. Ensuite, nous présenterons nos deux terrains d'investigation, les universités de Bordeaux 1 et Bordeaux 2, et la méthodologie de constitution de corpus d'ULH sur deux temporalités rapprochées, 2009 et 2011. Nous montrerons la représentativité des corpus obtenus sur un corpus aléatoire de sites web du domaine universitaire.

Enfin, nous présenterons quelques résultats issus de la comparaison quantitative de nos corpus d'ULH, pour en dégager la dynamique.

2 Description du corpus source des unités lexicales

Bien que comparable aux constructions de ressources terminologiques issues de corpus de textes ou d'entretiens d'experts (Rousselot & Frath, 2002), l'utilisation d'unités lexicales extraites des menus de navigation est particulière. Ces termes sont relativement faciles à repérer automatiquement mais ils ne sont pas « enchâssés dans le co-texte ». Ces termes sont par essence censés représenter les contenus profonds qu'ils relient, en quelque sorte des descripteurs

ou métadonnées de l'ensemble des contenus du site (Pinède & Reymond, 2011).

2.1 Site web organisationnel

Nous avons choisi de travailler sur des corpus de sites représentant un certain type d'organisation : nous appellerons ces sites « sites web organisationnels » (SWO). Ce choix de constitution de corpus, sous couvert d'organisations relevant d'un même domaine (au plan de leurs activités), garantit un facteur de cohérence et d'homogénéité, condition nécessaire au traitement terminologique (Bourigault et Charlet, 2004).

2.2 Page d'accueil et ULH

La page d'accueil présente l'information gérée par l'organisation ou la structure concernée, tout en traduisant des choix sélectifs et stratégiques (Nielsen, Tahir, 2002) par les ULH.
Dans le cadre de notre recherche, nous considérons donc la page d'accueil en tant que « sommaire » du site (au plan lexical), qui joue le rôle de menu pour le reste du site (Piotrowski 2009).
Nous avons montré que les ULH de la page d'accueil, produisent une signature textuelle et sémantique de cette page (Reymond et Pinède, 2010a).
Après avoir montré la représentativité de la taxonomie constituée, nous nous intéressons ici à sa dynamique temporelle, par la comparaison de deux corpus d'ULH issus d'un ensemble de SWO du domaine universitaire. Nous montrons la tendance globale de cette dynamique à court terme par une comparaison diachronique établie entre 2009 et 2011. L'intérêt de ces travaux est de montrer la stabilité de ces termes au sein d'un même domaine, malgré la dynamique générale du web.

3 Méthodologie

3.1 Constitution des corpus d'ULH

Nous avons collecté les ULH présentes sur les interfaces de pages d'accueil de sites web relevant du domaine organisationnel de deux établissements universitaires de sciences (Bordeaux 1) et de santé (Bordeaux 2). Nous montrons la représentativité de cet ensemble terminologique pour évaluer sa transférabilité immédiate sur d'autres contextes universitaires. Nous définissons le taux de recouvrement comme le nombre d'ULH présentes et faisant partie du corpus par rapport au nombre total d'ULH sur une page. Nous comparons ce critère sur les sites en rapport avec les sites de la collecte terminologique (Tableau 1) puis sur des sites de zones DNS d'établissements de villes aléatoirement sélectionnées.
L'analyse de la dynamique du corpus nous contraint à réduire par sélection les sites présents simultanément en 2011 et sur le site webarchive.org pour la collecte 2009.
Au final, notre corpus de site pour la comparaison est constitué de 96 pages d'accueil, 48 issues des domaines DNS de l'université de Bordeaux 1, et 48 des DNS de Bordeaux 2. Les ULH ainsi recueillies ont été classées dans une taxonomie de représentation de la communication web de ces organisations. Le lecteur intéressé suivra une présentation de cette construction (Pinède et Reymond, 2010,2011) ainsi que résultats de quelques tests applicatifs (Reymond & Pinède, 2010a, 2010b), ainsi qu'un approfondissement de cette étude de dynamique terminologique sur les classes de la taxonomie (Pinède et. al, 2001).
Pour chaque année (2009 et 2011) nous mesurons le nombre total d'ULH (ULHt) ainsi que le nombre d'ULH distinctes (ULHd) par leur forme au plan lexical.

4 Résultats

4.1 Taux de recouvrement

Le tableau 1 montre la représentativité du vocabulaire de la taxonomie sur les deux universités de constitution. Quelques pages présentent un taux à 0 % : les pages d'accueil de ces sites sont en général réduites à leur plus simple expression, et ne sont donc pas représentatives.
Les deux zones DNS de référence n'obtiennent pas des taux de 100% du fait des ULH non pertinentes, ou encore de la présence de liens d'actualités, donc dynamiques et volatiles.

	Bx1	**Bx2**
Nombre de sites analysables	78	77
Tx à zéro	2	2
Recouvrement moyen	76 %	74 %
Recouvrement moyen hors zéros	78 %	75 %

Tableau 1 : Pouvoir de recouvrement de la taxonomie sur les universités Bordeaux 1, 2

Dans le Tableau 2 figurent les résultats obtenus sur un autre établissement bordelais (Bordeaux 3) relevant d'un autre secteur disciplinaire

(Lettres), et montrant un taux de couverture tout à fait comparable à ceux obtenus précédemment.

	Bx3	UPPA	Tlse1	Tlse2	Tlse3
Nbsites	41	105	26	97	106
Tx à 0	2	4	3	3	3
Tx moyen	70 %	70	61	74	69
Hors nuls	73 %	73	70	76	70

Tableau 2 : Pouvoir de recouvrement de la taxonomie sur les zones DNS des universités de Bordeaux 3, Pau et Pays de l'Adour, Toulouse 1, 2 et 3.

Le Tableau 2 montre encore ce taux sur l'UPPA, un établissement régional multidisciplinaire, et enfin sur les sites des universités de la ville de Toulouse (1, 2, 3).

	Montp.1	Montp.2	Montp.3	Nice
Nbsites	44	115	22	90
Tx à 0	3	6	4	11
Tx moyen	73	65	64	53
Hors nuls	79	70	79	61

Tableau 3 : Pouvoir de recouvrement de la taxonomie sur les zones DNS des universités de Montpellier 1, 2 et 3, et Nice.

De même (Tableau 3) les tests réalisés sur les universités de Montpellier et de Nice se révèlent extrêmement rassurants. Les taux de recouvrement obtenus ici sont tout à fait satisfaisants. L'écart pour l'université de Nice est dû à la présence d'un nombre important de pages d'accueil en anglais lors du recueil. Tout ceci valide notre hypothèse initiale quant à la représentativité du corpus d'ULH sur les web issus du domaine des universités françaises.
Les récurrences lexicales observées démontrent les régularités de structuration informationnelle et éditoriale dans les pages d'accueil des sites web de type universitaire.

4.2 Dynamique temporelle du corpus

Conformément aux attentes, une augmentation du nombre d'ULH s'affirme en 2011 par rapport à 2009. Le Tableau 4 montre les caractéristiques générales des collectes terminologiques sur les 96 sites.

	2009	2011
Nb d'ULH (avec occurrences) – ULHt	2096	2940
Nb d'ULH distinctes (ULHd)	1597	1895
Proportion des ULHd / ULHt	76 %	64 %

Tableau 4 : Caractéristiques des ULH collectées et retenues.

Toutefois, la ligne des ULH distinctes montre que cette augmentation de près de 40% n'est plus que de l'ordre de 18% au plan lexical.
Le tableau 5 montre la dynamique globale du corpus des ULH. Le sous ensemble 2009-2011 représente les ULH présents sur les deux années[1].

	ULHt	ULHd	Proportion
Disparus 2009	841	779	92,6 %
Communs 2009-2011	2762	818	29,6 %
Apparus 2011	1433	1077	75,2%
Total	5036	2674	53,1%

Tableau 5 : La dynamique du corpus des ULH. Les disparus, les communs aux deux années et les apparus.

Les spécifiques de 2009 sont des ULH qui ont disparues entre les deux périodes de collecte. Les spécifiques 2011 sont des ULH apparues en 2011. Trois remarques principales sont à mentionner :

- L'ensemble commun entre les deux corpus est de 53,2%, alors que seuls 29,6% sont communs sur le sous ensemble des ULH distinctes ;
- Le sous ensemble d'intersection 2009-2011, montre un principe d'occurrence très important (3 en moyenne) ;
- la comparaison des deux corpus spécifiques (2009 et 2011) montre en 2009 une très grande dispersion des ULH que l'on ne retrouve pas de façon si marquée en 2011 (plus que 75%).

5 Conclusion

Nous avons montré la représentativité des sous-ensembles terminologiques. La collecte sur deux zones éditoriales (plus de 100 sites), suffit pour obtenir un corpus représentatif des menus des pages d'accueil du domaine universitaire. Au plan de sa dynamique, le volume global des corpus d'ULH a nettement progressé entre 2009 et 2011. Les ratios montrent une convergence, autrement dit une normalisation terminologique, des ULH nouvellement utilisées vers un noyau

[1] Le nombre total ULHd dans ce découpage du corpus (2009-2011) est inférieur au nombre total d'ULHd du tableau 4 (1592+1891) du fait que dans ce dernier, la cardinalité de l'ensemble ULHd *commun 2009 et 2011* est comptée deux fois.

relativement stable (l'intersection des ensembles 2009 et 2011). De surcroît, les nouveaux termes apparaissent avec une multiplicité importante diminuant la proportion d'« hapax » de la zone éditoriale. D'autant que nous avons considéré pour cette analyse les diverses et nombreuses formes lexicales possibles. Ces résultats nous permettent d'apprécier la représentativité et la dynamique pour affiner notre approche. Nous envisageons d'abord de réaliser un traitement lexical (incluant les flexions, les marques de langage et lemmatisation) afin d'appuyer un système d'apprentissage automatique. Les perspectives sont de mettre en œuvre un classement en taxonomie (Rastier *et al.*, 1994, Tellier, 2009), dirigé au plan sémantique, tout en soulignant les singularités issues de la considération du degré méta-linguistique du vocabulaire traité. De plus, un des autres objectifs de ce travail est de pouvoir appliquer notre méthodologie à d'autres domaines (collectivités territoriales, sites commerciaux, etc.) afin d'étendre les domaines d'application de la taxonomie.

Références

Aussenac-Gilles N. et Condamines A. (2004), « Documents électroniques et constitutions de ressources terminologiques et ontologiques », in Information-*Interaction-Intelligence*, vol. 4, n°1, novembre, p 75-93.

Bourigault D., Charlet, « Construction de ressources terminologiques ou ontologiques à partir de textes : un cadre unificateur pour trois études de cas », *Revue d'Intelligence Artificielle (RIA)*. PIERREL J.M. et SLODZIAN M. (Ed.). Paris : Hermès. 18 (1) : 87–110. 2004.

Nielsen J., Tahir M. *L'art de la page d'accueil*. Paris, Eyrolles, 2002.

Otman, G. (dir), 1995, « Terminologie et intelligence artificielle », *La Banque des mots*, Numéro spécial, Paris, Conseil international de la langue française, n°7, 112 p.

(Pinède & Reymond, 2010) Pinède N., Reymond D. De la diversité au lissage informationnel : création d'une taxonomie inductive pour les sites web universitaires. *17e congrès de la SFSIC: au cœur et à la lisière des SIC*, Dijon 23-26 juin 2010.

(Pinède & Reymond, 2011) Pinède N., Reymond D. « Approche extensive des métadonnées pour un site web : principes d'élaboration et applications d'une taxonomie », *Etudes de Communication*, N°36, 2011.

(Pinède *et al.*, 2011) Pinède N., Reymond D., Lespinet-Najib V., Le Blanc B. « Terminologie hypertexte : dynamique temporelle d'une taxonomie », actes du 14e colloque international sur le document électronique, 7-8 décembre 2011, Rabat, (à paraître).

Piotrowski D., *L'hypertextualité ou la pratique formelle du sens*, Collection Lettres numériques, N°6, 2009.

Rastier, F. Cavazza, M, Abeillé, A. (1994) *Sémantique pour l'analyse : de la linguistique à l'informatique*, Paris, Masson.

Rastier F., Sémantique et recherches cognitives, PUF, 2010.

Reymond D. *Dynamique informationnelle d'une ressource Web: apport sémantique de la taxinomie. Étude webométrique des sites des universités françaises.* PhD thesis, Universités de Bordeaux - Michel de Montaigne. Dec. 2007.

Reymond D., Pinède N. "Using a taxonomy based fingerprint: classification and recognition of the academic webspace". *Proceedings of the Sixth International Conference on Webometrics, Informetrics and Scientometrics* (WIS) & Eleventh COLLNET Meeting, Mysore, India, 2010.

Reymond D., Pinède N. "Website and communication strategy alignment: a librarian science approach to webometrics tools". *Proceedings of the Sixth International Conference on Webometrics, Informetrics and Scientometrics* (WIS) & Eleventh COLLNET Meeting, Mysore, India, 2010.

(Rouquette, 2009) Rouquette S. *L'anayse des sites internet. Une radiographe du cyberspace*. Bruxelles, de Boeck , 2009.

(Rousselot & Frath, 2002) Rousselot, F. et Frath, P. (2002). « Terminologie et Intelligence Artificielle. Traits d'union. » éd. Kleiber, G. et Le Querler, N. p. 181-192. Presses Universitaires de Caen, Caen.

Tellier I., « Apprentissage automatique pour le TAL : Préface », *Traitement Automatique des Langues* 50, 3 (2009) p.7-21.