Identifier des traces d'innovation : Proposition d'une approche outillée en corpus spécialisés

Aurélie Picton[1]

[1]Observatoire de linguistique Sens-Texte
Université de Montréal
C.P. 6128, succ. Centre-ville
Montréal (Québec), Canada H3C 3J7
aurelie.picton@umontreal.ca

Résumé : Dans cette communication, nous présentons une approche linguistique outillée en corpus pour repérer des traces d'innovation dans les textes spécialisés. Cette exploration, basée sur quatre indices linguistiques et une collaboration avec des experts de domaine, permet de dresser un portrait de l'innovation en terminologie avec un regard original et riche. Nos observations prennent appui sur l'exemple du domaine spatial.

Mots-clés : Diachronie, Langues de spécialité, Linguistique de corpus, Linguistique outillée, néologie.

1 Introduction

Les études diachroniques en terminologie restent aujourd'hui encore relativement rares. Or, l'évolution des connaissances est un phénomène central et inévitable dans les domaines scientifiques et techniques. Dans le domaine spatial, la question de l'évolution prend une dimension particulière dans le cadre de la mise en place de projets spatiaux qui s'étendent généralement sur une dizaine d'années et impliquent d'observer l'évolution en « diachronie courte ». Pour ce faire, une des possibilités est d'apprendre à identifier comment ces évolutions – et plus particulièrement dans cette présentation l'innovation - peuvent être repérées dans les textes.

Ce type de démarche ouvre la voie à la réflexion sur la prise en compte de la dimension diachronique en terminologie textuelle, en interrogeant notamment les méthodes d'exploration de l'évolution en corpus, le rôle de l'expert dans la tâche d'analyse et le lien entre textes et connaissances en diachronie. Nous présentons ici la démarche mise en œuvre et insistons sur la nature de l'innovation en jeu sur de courts intervalles temporels que cette approche permet de mettre au jour.

2 Hypothèse

La démarche d'exploration de corpus spécialisés en diachronie peut reposer sur l'hypothèse classique en terminologie textuelle selon laquelle les connaissances partagées par les experts de domaine sont accessibles dans les textes (Bourigault & Slodzian, 1999 : 30). Plus précisément, comme le formule Cabré (1998 : 141), l'hypothèse que l'on peut poser est que : « [l]es variations de sens/de concepts peuvent être mises au jour à partir du repérage de variations d'expression dans les textes du domaine ». Autrement dit, il est possible de repérer en corpus des variations linguistiques que l'on peut interpréter et associer à des évolutions de connaissance dans le domaine. Nous nous concentrons ici sur le repérage de l'innovation et proposons quatre indices linguistiques pour l'identifier. Dans ce cadre, notre approche se distingue des démarches qui visent l'extraction de néologismes seuls et qui se basent sur la comparaison de corpus et de listes d'exclusion pour repérer ces néologismes (par exemple Cabré, et al., 2003 ; Roche & Bowker, 1999 ; Janssen, 2008).

3 Méthodologie

3.1 Deux corpus : TTVS et DORIS

Pour correspondre au cadre de la diachronie courte imposée par notre contexte d'étude, deux corpus ont été construits. Le premier est constitué des chapitres d'optique et d'optoélectronique spatiale du cours de Techniques et Technologie des Véhicules Spatiaux (désormais TTVS), édité tous les 4 ans depuis 1994 par le CNES aux Éditions Cépaduès (Cnes, 1994, 1998, 2002). Ce cours est rédigé par plus de 80 experts du CNES, à l'attention de semi-experts (Bowker & Pearson, 2002) et contient une dizaine de chapitres qui englobent l'ensemble des domaines de compétence du CNES.

Table 1. Nombre d'occurrences dans le corpus TTVS

	TTVS1994	**TTVS1998**	**TTVS2002**	**Total**
Nombre d'occurrences	46 448	78 656	109 505	234 609

Le second corpus, DORIS, est un corpus « projet » constitué des rapports de spécification des première et troisième générations de balises DORIS. DORIS (Détermination d'Orbite et Radiopositionnement Intégrés par Satellite) est un système de positionnement de satellites par balises terrestres dont la première génération a été conçue et développée à la fin des années 80 et la troisième génération au début des années 2000. Les textes sont donc répartis en deux sous-corpus : génération 1 (DORISGEN1) et génération 3 (DORISGEN3).

Table 2. Nombre d'occurrences dans le corpus DORIS

	DORISGEN1	**DORISGEN3**	**Total**
Nombre d'occurrences	17 544	18 857	36 401

3.2 Démarche outillée

La démarche d'exploration proposée s'appuie sur l'utilisation d'outils classiques d'exploration de corpus tels qu'un concordancier (AntConc, Anthony, 2005), un extracteur de termes (Syntex, Bourigault, et al., 2005), un extracteur de relations (TerminoWeb, Barrière & Agbago, 2006). Ce choix méthodologique inscrit notre démarche dans la lignée des travaux en terminologie textuelle, notamment pour la création de ressources termino-ontologiques.

3.3 Collaboration avec des experts

Les outils choisis permettent d'extraire des termes du domaine et des données sur des variations linguistiques entre les différentes générations des corpus, à l'aide des indices décrits *infra*. Ces résultats sont présentés et discutés avec des experts du domaine pour garantir la qualité de l'interprétation de ces indices dans notre perspective diachronique.

3.4 Quatre indices linguistiques

À ce jour, nous avons travaillé à partir de quatre indices linguistiques associables à des traces d'innovation dans les textes :

- Les empreintes de fréquence : cet indice se concentre sur l'observation des termes qui apparaissent au fil du temps dans les corpus.

Table 3. Empreinte de fréquence d'un terme/concept nouveau (*APS*)

	TTVS1994	**TTVS1998**	**TTVS2002**
APS	0	2	72

Dans l'exemple ci-dessus, l'empreinte de fréquence du terme/concept *APS* révèle qu'il n'apparaît dans le corpus qu'à partir de 1998 et massivement à partir de 2002. On peut donc émettre l'hypothèse qu'il s'agit d'un candidat néologisme.

- L'identification de contextes riches en connaissances évolutives : il s'agit de repérer des portions de textes qui contiennent des informations pertinentes quant à l'évolution du domaine. Ces informations sont repérables à l'aide de marqueurs tels que des adverbes (*aujourd'hui*, *récemment*), des adjectifs (*nouveau*, *prometteur*), des noms (*prototype*, *innovation*), des verbes (*inventer*). Une fois ces marqueurs définis, les termes qui apparaissent dans ces contextes sont collectés.

Figure 1. Exemple de contexte riche en connaissances évolutives

*Un produit **nouveau** est **apparu depuis quelques années** sur le marché, il s'agit des multi-barrettes.*

Dans cet exemple, le terme/concept *multi-barrettes* peut lui aussi être considéré comme un candidat néologisme.

- La coexistence de variantes de termes : l'apparition de concepts nouveaux peut être accompagnée par la création de plusieurs dénominations concurrentes pour ces concepts (Dury & Lervad, 2007 ; Guilbert, 1965). Dans ce cas, les variantes peuvent être considérées comme des indices pertinents d'innovation.

Table 4. Variantes concurrentes (*visibilité de satellite* et *visibilité satellite*)

	DorisGen1	**DorisGen3**
visibilité de satellite	0	4
visibilité satellite	0	8

L'exemple ci-dessus présente un cas où deux dénominations équivalentes et concurrentes apparaissent simultanément en corpus. Ce phénomène de « concurrence » entre dénomination peut indiquer la nouveauté du terme/concept en question (Guilbert, 1975 ; Dury, 2008). De la même manière, comme illustré dans la Table 5, lorsque plusieurs dénominations existent, au fil du temps, l'une d'entre elle peut se stabiliser et s'implanter au détriment des autres. Dans l'exemple de *synthèse d'ouverture*, le terme/concept est nouveau en 1994 et s'implante de plus en plus dans le domaine, comme en témoignent les fréquences.

Table 5. Variantes concurrentes et implantation (*synthèse d'ouverture*)

	TTVS1994	**TTVS1998**	**TTVS2002**
synthèse d'ouverture optique	5	1	0

synthèse d'ouverture	4	3	7
SO	3	0	0

- Les dépendances syntaxiques : comme décrit par Ahmad et al. (2002) « syntactical productivity is quite apparent in specialized texts. When specialists write about a concept or an artefact, they start by describing one concept or artefact. Soon enough they find concepts or artefacts which they can relate to their original and, indeed, often form classes of concepts or artefacts ». Par conséquent, la spécification d'un concept existant peut être considérée comme un indice pertinent d'innovation. C'est ce qu'illustre l'exemple de la Table 6.

Table 6. Création d'une nouvelle dépendance syntaxique (*circuit hybride*)

Terme	**Descendant**	**TTVS1994**	**TTVS1998**	**TTVS2002**
circuit		30	42	53
	circuit hybride	0	12	13

De la même manière, un changement de dépendances peut également être associé à de la nouveauté (Table 7) :

Table 7. Changement de dépendances (*opérateur*)

Terme	**Descendant**	**DorisGen1**	**DorisGen3**
opérateur		89	26
	interface opérateur	0	12
	opérateur appuyer sur touche	26	0
	opérateur appuyer sur val	5	0
	opérateur tourner clé	6	0

Dans cet exemple, le descendant *interface opérateur* renvoie à un élément nouveau sur la balise DORIS, nouveauté qui laisse également entrevoir une innovation dans le rôle de l'opérateur des balises: toutes les dépendances verbales où le verbe implique une opération manuelle dont l'opérateur est l'agent disparaissent et laisse affirmer que le rôle de l'opérateur bénéficie aujourd'hui de plus en plus de support informatique dans ses actions (*interface*).

4 Portrait de l'innovation dans le domaine spatial

Généralement, l'innovation dans les domaines de spécialité est essentiellement associée à l'apparition de néologismes. Or, l'analyse des quatre indices proposés *supra* permet d'affiner la description de ce que peut être l'innovation dans un domaine de spécialité en diachronie courte, innovation repérable en corpus. La description peut être résumée ici en trois points saillants.

4.1 Relativité de la nouveauté et nature du domaine observé

L'innovation repérable en corpus peut être mise en lien avec les notions de « domaine de connaissance » et « domaine d'activité» proposées par De Bessé (2000). Bien que critiquables, ces définitions permettent de proposer une distinction intéressante sur la nouveauté : dans le premier cas, celui d'un domaine de connaissance (représenté par le corpus TTVS), l'innovation observée entre dans le champ d'un « savoir constitué, structuré, systématisé selon une thématique » (*ibid*.). C'est dans ce cadre que la terminologie aborde généralement la notion de néologie. Néanmoins, dans un cas des projets spatiaux, comme dans le corpus DORIS, il ne s'agit plus d'un domaine de connaissance, mais plutôt un domaine d'activité, qui « correspond à une activité humaine […] constitué[e] d'un ensemble de procédés bien définis destinés à produire certains résultats » (*ibid*.). Dans ce cadre, l'innovation n'est plus associable à de la néologie. Par exemple, dans le corpus DORIS, le terme/concept *visibilité satellite* est nouveau dans le projet DORIS, mais ne l'est pas dans le domaine spatial. La nouveauté de ce concept est donc uniquement relative à un projet donné. Il est donc difficile de parler de néologie au sens « classique ».

4.2 Nécessité de distinguer dénomination, concept et instance en diachronie

Le deuxième point que l'exploration proposée permet de souligner concerne l'importance de dissocier plusieurs niveaux d'évolution. L'innovation peut en effet concerner :

- les dénominations seules (néologie de forme),
- le concept seul (néologie de sens),
- la dénomination et le concept ensemble (néologie « totale »)
- mais aussi celui de l'instance de concept.

Ce dernier point est rarement souligné, mais concerne les cas où la nouveauté intervient sur une instance particulière d'un concept, qui lui n'est pas nouveau. Par exemple, dans le cas du corpus DORIS, la forme de la balise est susceptible de changer au fil des générations successives. Un interrupteur peut apparaître sur le

boîtier, une touche sur le clavier, etc. Dans ces cas, les termes/concepts *touche* ou *interrupteur* ne sont pas nouveaux. Ce sont la touche spécifique ou l'interrupteur spécifique qui apparaissent sur la balise qui sont nouveaux. Dans le corpus TTVS, on peut prendre l'exemple du terme/concept *Modèle instrumental* : dans le cadre du projet SPOT, un modèle instrumental spécifique nouveau a été défini. Le terme/concept n'est donc pas nouveau en lui-même, mais le modèle instrumental spécifique défini pour SPOT à ce moment-là l'est.

4.3 Vers le repérage d'autres types d'innovation

Enfin, la démarche proposée permet également de mettre au jour des aspects d'innovation dans les connaissances du domaine tels que des améliorations de techniques ou concepts existants, des progrès (outils plus performants, composants plus légers, moins encombrants, etc.), des modifications (de la structure des documents par exemple), etc. repérables linguistiquement en corpus. De fait, cette démarche permet d'envisager non plus le repérage de néologismes seuls, mais plutôt le repérage d'informations multiples sur l'évolution des domaines, informations susceptibles d'intéresser de nombreuses applications telles que la mise à jour de ressources terminologiques, la veille scientifique et technique, la recherche d'information, etc. Ce type de repérage ouvre donc de nombreuses perspectives à creuser dans cette voie. Ainsi, dans nos travaux de thèse (Picton, 2009), 17 types d'évolution ont pu être mis au jour et participer ainsi à une première typologie de l'évolution en diachronie courte.

5 Conclusion et perspectives

La démarche présentée dans cette communication repose sur l'analyse outillée d'indices linguistiques et sur la collaboration avec des experts du domaine. L'un des intérêts de ce type d'approche est de permettre de faire émerger des aspects de l'innovation très peu observés jusqu'à aujourd'hui mais présents dans les domaines de spécialité. Ces possibilités nouvelles de description permettent de mieux comprendre ce que peut être l'innovation dans les domaines de spécialité et présente un potentiel intéressant pour de nombreuses applications telles que la mise à jour de ressources terminographiques, la veille scientifique, etc. De la même manière, cette démarche, en plus de l'apport descriptif qu'elle offre, permet de poser certaines questions pour mieux baliser la prise en compte de la dimension diachronique en terminologie textuelle et en particulier les questions de l'outillage et des possibilités d'exploration de corpus spécialisés diachroniques, la question de l'interprétation en corpus et de la collaboration entre linguistes/terminologues et experts en diachronie, et enfin la question du lien entre langue et connaissances.

Références

AHMAD, K., SCHIERZ, A. & AL-THUBAITY, A. (2002). Discovery and Terminology. In Actes de la conférence internationale "Terminology and Knowledge Engineering" (TKE 2002), Nancy, France, 28-30 août 2002, p.1-6.

ANTHONY, L. (2005). AntConc: Design and Development of a Freeware Corpus Analysis Toolkit for the Technical Writing Classroom. In Actes de la conférence "Professional Communication Conference" (IPCC 2005), Limerick, Irlande, 13 juillet 2005, p.729-737.

BARRIÈRE, C. & AGBAGO, A. (2006). TerminoWeb: A Software Environment for Term Study in Rich Contexts. In Actes de la conférence internationale "Terminology, Standardisation and Technology Transfer" (TSTT 2006), Beijing, Chine, 25-26 août 2006, p.103-113.

BESSE (DE), B. (2000). Le domaine. In H. BEJOINT & P. THOIRON Éds. *Le sens en terminologie*. Presses Universitaires de Lyon, Travaux du CRTT (Centre de Recherche en Terminologie et Traduction), Lyon, p.182-197.

BOURIGAULT, D., FABRE, C., FREROT, C., JACQUES, M.-P. & OZDOWSKA, S. (2005). Syntex, analyseur syntaxique de corpus. In Actes de la 12ème conférence "Traitement Automatique des Langues Naturelles" (TALN 2005), Dourdan, France, 6-10 juin 2005, p.17-25.

BOURIGAULT, D. & SLODZIAN, M. (1999). Pour une terminologie textuelle. *Terminologies Nouvelles*, 19, p.29-32.

BOWKER, L. & PEARSON, J. (2002). *Working with Specialized Language: a Practical Guide to Using Corpora*. Routledge, London/New York.

CABRÉ, M. T. (1998). *La terminologie : théories, méthodes et applications*. Armand Colin, Presses de l'Université d'Ottawa, Ottawa.

CABRE, M. T., DOMENECH, M., ESTOPA, R., FREIXA, J. & SOLE, É. (2003). L'observatoire de néologie : conception, méthodologie, résultats et nouveaux travaux. In J. F. SABLAYROLLES Éd. *L'innovation Lexicale.* Honoré Champion, Paris, p.125-147.

CNES (1994). *Cours de Techniques et Technologies des Véhicules Spatiaux (3 volumes)*. Centre National d'Études Spatiales, Toulouse.

DURY, P. (2008). Les noms du pétrole : une approche diachronique de la métonymie onomastique. *Lexis , « Polysemy / La polysémie »*, 1, p.10-22.

DURY, P. & LERVAD, S. (2007). La variation dans la terminologie de l'énergie : approches synchronique et diachronique, deux études de cas. In Actes du colloque "Terminologie : approches transdisciplinaires" (communication orale), Gatineau, Québec, Canada.

GUILBERT, L. (1965). *La formation du vocabulaire de l'aviation*. Thèse de Doctorat ès Lettres, Faculté des Lettres et Sciences Humaines, Université de Paris, Larousse, Paris.

GUILBERT, L. (1975). *La créativité Lexicale*. Larousse, Paris.

JANSSEN, M. (2008). NeoTrack – Un analyseur de néologismes en ligne. In Actes du Ier Congrès International de Néologie des langues romanes (Cinéo 2008), Barcelone, Espagne, 07-10 mai 2008.

Picton, A. (2009). *Diachronie en langue de spécialité. Définition d'une méthode linguistique outillée pour repérer l'évolution des connaissances en corpus. Un exemple appliqué au domaine spatial*. Thèse de Doctorat en Sciences du Langage. Université Toulouse 2.

ROCHE, S. & BOWKER, L. (1999). Cenit : Système de détection semi-automatique des néologismes. *Terminologies Nouvelles*, 20, p.12-16.