Variation terminologique et analyse diachronique

Annie Tartier

LINA FRE CNRS 2729
annie.tartier@univ-nantes.fr

Résumé

Cet article présente un travail destiné à automatiser l'étude de l'évolution terminologique à partir de termes datés extraits de corpus diachroniques de textes scientifiques ou techniques. Les apparitions et disparitions d'attestations de termes au cours du temps constituent la manifestation la plus simple de l'évolution. Mais la prise en compte des formes variantes apporte une information de meilleure qualité sur le suivi des termes. Une distance entre termes complexes permet de rendre opérationnelle l'intégration de la variation terminologique à l'analyse diachronique. Des résultats montrant la prise en compte des variantes sont présentés et commentés à la fin de l'article.

Mots-clés : évolution terminologique, corpus diachronique, terme complexe, variation terminologique, distance.

Abstract

This paper presents a work about terminological evolution of dated terms, extracted from diachronic corpora of scientific or technical texts. Disappearance and appearances of terms attestations are the most obvious evidence of change. Nonetheless taking into account the variant forms of terms helps describe their historic development. A distance between complex terms allows terminological variation to be integrated into diachronic analysis. Some results and comments are given at the end of the paper.

Keywords: terminological evolution, diachronic corpus, complex term, terminological variation, distance.

1. Introduction

Nous ouvrons cet article par deux citations destinées à convaincre, s'il en était besoin, que les langues évoluent sans que l'homme en ait conscience, que les scientifiques ont besoin de maîtriser cette évolution et qu'ils n'ont que très peu d'outils pour le faire.

> « Le *théoricien* (scientifique, chercheur) a besoin d'être tenu au courant de l'évolution récente de son domaine, et donc de maîtriser les notions et les termes nouveaux ... »
>
> « ... il suffira de mentionner l'incapacité des dictionnaires généraux à maîtriser les vocabulaires techno-scientifiques. » (Rey, 1992)

Le propos de cet article est de présenter, en réponse à ces besoins, des méthodes élaborées pour rendre compte de l'évolution des terminologies, de manière objective, avec peu de connaissances extérieures, en explorant automatiquement des corpus diachroniques de textes techniques ou scientifiques. Les objets d'observation sont les formes linguistiques de termes de la langue française. La volonté de rendre compte de manière objective interdit l'introspection, souvent pratiquée en linguistique, et conduit à travailler sur des données attestées, ayant valeur de témoignages. Il s'agit en particulier de suivre au cours du temps ce que devient un terme, et

quels sont les termes stables ou instables. Nous verrons que la maîtrise du phénomène de la variation terminologique prend ici toute son importance, parce que l'attestation d'un terme peut être repérée, même si celui-ci apparaît sous une forme variante.

Après avoir donné les idées fortes pour une analyse diachronique automatique, nous mettons l'accent sur la prise en compte du phénomène de la variation terminologique et son intégration au sein de l'analyse. Les méthodes présentées sont implémentées dans un prototype qui a permis de calculer les résultats affichés à la fin de l'article.

2. Cadre de l'étude

2.1. Études diachroniques, un panorama rapide

L'homme maîtrise mal les modifications : chacun a sans doute regretté de ne pas avoir pris de repère dans un état antérieur au moment où il se demande si ce qu'il observe dans l'état actuel est une nouveauté. La prise de conscience des disparitions est encore plus aléatoire et se fait souvent à l'occasion d'un événement fortuit. Enfin, lorsque les choses évoluent lentement on n'en perçoit pas le changement. Celui-ci ne sera rendu certain que si l'on pose des repères pour faire des comparaisons systématiques. Les outils informatiques mis à notre disposition nous offrent la possibilité d'effectuer ces repérages sur des données de grande taille.

Les travaux sur l'évolution de la langue se retrouvent dans au moins trois domaines qui se distinguent par leur objectif. En premier lieu, la linguistique historique, puis la linguistique « diachronique », mise en place par (De Saussure, 1916), s'appliquent à la langue, étudiée pour elle-même. Le deuxième domaine, la terminologie, accorde entre autre une part importante au recensement et à l'étude des néologismes. La veille scientifique, enfin, n'étudie l'évolution des terminologies que pour trouver des marques d'évolution des concepts portés par les termes. Il est intéressant de constater que si les objectifs de ces trois domaines n'ont rien de commun, les méthodes utilisées sont très proches. Les études d'accroissement lexical faites sur des textes littéraires (fonctionnalités proposées par le logiciel *THIEF*[1]) et les programmes de bibliométrie (Callon *et al*., 1993) qui analysent les documents techniques ou les textes des brevets sont tous basés, aux départ, sur les principes de la statistique lexicale (Muller, 1992).

L'un des points faibles relevés dans ces travaux est que seules les formes brutes des mots simples sont prises en compte. Ni les unités lexicales complexes, ni les variantes ne le sont, ce qui veut dire qu'un terme peut sembler avoir disparu alors qu'il est toujours attesté sous une forme variante. C'est un défi que nous avons voulu relever en introduisant la variation terminologique au sein d'une analyse diachronique.

2.2. Démarche générale

Étudier l'évolution d'un système revient à croiser l'observation de ce qui change et le repérage de ces changements sur un axe du temps. Le matériau d'étude est constitué d'ensembles de formes terminologiques datées.

Une partie de l'étude concerne l'observation de ce qui change. Les apparitions et disparitions d'attestations de certains termes au cours du temps constituent la manifestation la plus simple du changement. Elles se mesurent par une information de type présence/absence. Mais une étude plus fine montre que la maîtrise du phénomène permet de regrouper les variantes d'un

[1] http://ancilla.unice.fr/\~{}brunet/pub/THIEF/THIEF1.html

même terme et de ne pas les considérer comme des termes différents dans l'observation des phénomènes d'évolution. D'autre part certains auteurs, comme (Cabré, 1998), avancent l'hypothèse que l'évolution, globale, se ferait pas par le moyen de variations locales, certains nouveaux termes résultant de la lexicalisation de termes existants. En conséquence, mettre à jour des variations pourrait permettre de révéler une évolution. Pour contrôler la variation de manière opérationnelle nous avons introduit une distance entre deux formes terminologiques. Sa mise en place est développée dans la section 3.

La section 4 explique comment la prise en compte de la variation terminologique est intégrée aux calculs qui visent à révéler l'évolution des attestations de termes.

3. Variation terminologique et distance

3.1. Variation terminologique : une typologie

Il existe de nombreuses typologies de la variation (Jacquemin, 1999 ; Daille, 2002), conçues le plus souvent en fonction des objectifs de leurs auteurs. Nous avons retenu la typologie suivante :

- variation orthographique (*quasiélastique* / *quasi-élastique*) ;
- variation morphologique : dérivation, composition (*photoélectron* / *électron*) ;
- variation syntaxique qui se décompose en :
  - syntaxique faible (*diffusion laser* / *diffusion par laser*) ;
  - présence ou ellipse d'un nom support (*diffraction électronique rasante* / *diffraction électronique en incidence rasante*) ;
  - insertion, expansion, substitution (*diffraction de neutrons* / *diffraction élastique de neutrons*) ;
- variation morphosyntaxique (*diffraction électronique* / *diffraction des électrons*) ;
- variation « sémantique » : synonymie, hyperonymie ;
- sigles et abréviations.

3.2. Distance entre deux formes terminologiques

La variation entre deux formes terminologiques a comme effet un glissement de sens, plus ou moins important, que l'on tente de quantifier par une *distance*. Il est alors aisé de considérer comme variantes des formes dont la distance est inférieure à un seuil, paramétrable selon l'application souhaitée. Il ne peut bien sûr être question de réduire un phénomène linguistique à un nombre. Mais la disponibilité d'une telle mesure permet aux programmes de regrouper des formes variantes en s'appuyant sur un seuil défini au préalable.

La distance informationnelle (Bennett *et al.*, 1998) définit la distance entre deux entités complexes par la longueur du plus court programme permettant de décrire la transformation d'une entité dans l'autre. Cette distance s'adapte bien aux termes complexes pour lesquels on peut définir des transformations correspondant aux différents types de variations. Pour la rendre opératoire, nous avons repris la distance d'édition entre chaînes de caractères (Levenshtein, 1966), qui est justement fondée sur le principe de la transformation.

Après avoir brièvement rappelé le principe de calcul de cette distance, nous expliquons comment nous l'avons adapté à la distance entre deux termes complexes.

3.3. Distance d'édition entre chaînes

Il s'agit de transformer une chaîne en une autre par des applications de trois opérations élémentaires sur certains caractères (insertion, suppression, substitution). Chaque transformation globale, résultant des opérations élémentaires, se traduit par un alignement. Un coût est attribué à chaque opération élémentaire. La coût d'un alignement est égal à la somme des coûts des opérations élémentaires impliquées et le coût minimal constitue la distance recherchée.

Il est important de remarquer que la distance est indépendante de la nature et de la position des caractères auxquels s'appliquent les opérations.

Les alignements de coût minimum et la distance correspondante sont calculés par un algorithme de programmation dynamique, initialement attribué à (Wagner et Fischer, 1974).

3.4. Adaptation aux termes complexes

L'idée consiste à faire jouer aux constituants des termes, le rôle que jouent les caractères dans la distance d'édition. C'est ainsi que l'ont peut passer d'une forme à une autre en alignant au mieux certains constituants.

Exemple :

diffusion	*cohérente*	*inélastique*	*d'électrons*	
diffusion	*cohérente*		*des neutrons*	*thermiques*
		1 suppression	1 substitution	1 insertion

Mais cette adaptation ne peut être une simple transposition. Pour réaliser cette adaptation, trois attributs sont associés à chaque forme d'un terme complexe :

– la suite des lemmes de chaque constituant dans laquelle les mots grammaticaux ont été regroupés avec les mots pleins qu'ils gouvernent ;
– la suite des étiquettes grammaticales constituant le schéma morphosyntaxique de la forme linguistique du terme (étiquettes de (Brill, 1992) : SBC (substantif commun), ADJ (adjectif), PREP (préposition), etc.) ;
– la suite des niveaux de dépendance : on donne 0 à la tête du terme, puis 1 à tous les éléments qui dépendent de cette tête, puis $n + 1$ à tous les éléments qui dépendent d'un élément de niveau n.

Exemple :

terme	: *diffusion inélastique de neutrons thermiques*			
lemme	: diffusion	inélastique	de_neutron	thermique
schéma	: SBC	ADJ	PREP_SBC	ADJ
niveaux	: 0	1	1	2

L'algorithme qui calcule la distance entre deux termes complexes est le même que celui de la distance d'édition entre chaînes. Mais, au lieu d'affecter un coût standard à chaque opération élémentaire, le coût est calculé par des règles qui intègrent le type de variation, en s'appuyant sur des *coûts de base*, paramétrables. Ces règles s'appuient sur la nature grammaticale des éléments, ainsi que, dans le cas de substitutions, sur l'existence éventuelle de liens morphologiques ou morphosyntaxiques. Par exemple, la substitution de *du neutron* à *neutronique* sera moins coûteuse que celle de *du neutron* à *électronique*. Plus l'élément concerné est « éloigné » de la tête du terme, moins l'opération doit coûter, c'est pourquoi les coûts des opérations sont pondérés par les niveaux de dépendance des éléments concernés. L'alignement d'éléments qui n'ont pas le même niveau de dépendance est interdit. Il arrive enfin qu'une variation syntaxique et une variation morphologique s'appliquent simultanément à la même paire d'éléments (*de_photoélectron / du_photo-électron*). Dans ce cas les « coûts morphologiques » sont ajoutés aux « coûts syntaxiques ».

3.5. Exemples

Les distances obtenues, présentées dans l'exemple ci-dessous, n'ont aucune valeur intrinsèque. Elles ne servent qu'à caractériser la situation relative de deux formes terminologiques en fonction des valeurs de seuil choisies.

diffusion de_neutron, diffusion neutronique	0.5	formes variantes
diffusion de_neutron, diffraction de_neutron	7.5	formes étrangères
diffusion neutronique, diffusion élastique de_neutron	1.5	formes parentes
diffusion lent de_neutron, diffusion du_neutron lent	1.75	formes parentes

4. Intégration dans l'analyse diachronique

Dans cet article nous avons délibérément pris le parti de ne pas détailler l'analyse diachronique. Nous ne présentons ici que ce qui est nécessaire pour comprendre comment la variation terminologique y est intégrée.

4.1. Segmentation par le temps

Le temps n'est qu'un repère qui permet de définir une relation d'ordre total (chronologique) entre des événements. On utilise un modèle à temps discret dans lequel l'axe est une suite d'intervalles élémentaires consécutifs et disjoints, marqués par un entier nommé *date*. Le temps est vu comme un ensemble de *chronons* munis d'une relation d'ordre (Fauvet et Scholl, 1995). Deux événements qui se projettent dans le même intervalle élémentaire sont supposés avoir lieu à la même date. Un intervalle quelconque est la réunion d'intervalles élémentaires consécutifs.

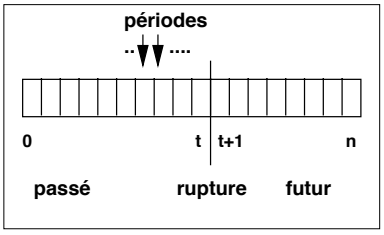

Figure 1. Vue temporelle

Les données de départ sont constituées d'un ensemble de formes terminologiques datées. Pour structurer ces données dans le temps on définit une *période* comme un ensemble de formes dont les dates d'attestations appartiennent à un intervalle. On peut donc appliquer aux périodes toutes les opérations sur les ensembles. On définit une *vue temporelle* comme la réunion d'une suite de périodes consécutives, concernées par une étude particulière. L'une des périodes $\mathcal{P}_t$,désignée comme la période de rupture, définit un *passé* et un *futur* (voir figure 1).

4.2. Événements sans prise en compte de la variation

Changement et évolution ne sont pas nécessairement synonymes. Certains changements sont éphémères. Il n'y a évolution que lorsque le changement est confirmé. Mesurer l'évolution ne consiste donc pas uniquement à repérer des occurrences de changements mais à s'assurer de leur durabilité. Les événements (stabilité, disparition, apparition) sur les formes exactes se calculent donc par des unions ($\bigcup$), intersections ($\bigcap$) et différences ($\setminus$) d'ensembles.

attestations stables	attestations disparues	attestations apparues
$\bigcap_{i=0}^{n} \mathcal{P}_i$	$\left(\bigcap_{i=0}^{t} \mathcal{P}_i\right) \setminus \left(\bigcup_{k=t+1}^{n} \mathcal{P}_k\right)$	$\left(\bigcap_{k=t+1}^{n} \mathcal{P}_k\right) \setminus \left(\bigcup_{i=0}^{t} \mathcal{P}_i\right)$

4.3. Extension aux variantes

Nous définissons une opération spécifique qui calcule l'ensemble des variantes d'une période dans une autre. Soient $\mathcal{P}_i$ et $\mathcal{P}_k$ deux périodes. Il existe, dans la différence $\mathcal{P}_k \setminus \mathcal{P}_i$ des formes f_k qui sont des variantes de formes f_i de $\mathcal{P}_i$. L'ensemble de ces formes constitue $\mathcal{V}_{\mathcal{P}_i}^{\mathcal{P}_k}$, l'ensemble des variantes de $\mathcal{P}_i$ dans $\mathcal{P}_k$. La distance, définie en 3.2, est utilisée par une fonction $variantes(f_k, f_i)$ pour comparer à un seuil la distance entre les deux formes f_k et f_i, et décider si celles-ci sont des variantes.

$$\mathcal{V}_{\mathcal{P}_i}^{\mathcal{P}_k} = \{f_k \in \mathcal{P}_k \setminus \mathcal{P}_i \ / \ \exists f_i \in \mathcal{P}_i \text{ avec } variantes(f_k, f_i)\}$$

La figure 2 illustre cette définition.

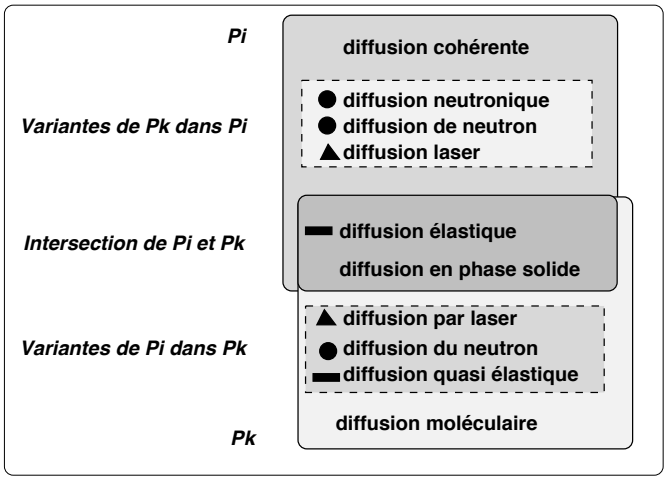

Figure 2. Périodes étendues aux variantes

Il suffit alors de remplacer dans les formules précédentes chaque période $\mathcal{P}$ par cette même période étendue à ses variantes dans la vue temporelle $\mathcal{P} \cup \mathcal{V}_P^{Vue}$.

Les définitions des événements deviennent alors les suivantes :
- un terme est dit *stable* si l'une de ses formes variantes est attestée dans chaque période de la vue temporelle ;
- un terme est dit *obsolète* si l'une de ses formes variantes est attestée dans chaque période du passé, et qu'aucune de ses formes variantes ne peut être trouvée dans le futur ;
- un terme est dit *nouveau* si aucune de ses formes variantes ne peut être trouvée dans le passé, et que l'une de ses formes variantes est attestée dans chaque période du futur.

5. Résultats

5.1. Le prototype

Les résultats présentés dans cette section ont été calculés grâce à certaines fonctions d'un prototype destiné à analyser l'évolution des terminologies (Tartier, 2004). Celui-ci est composé :

– de modules de prétraitement, dédiés à chaque source de données, qui renvoient une liste de formes terminologiques datées caractérisées par un quadruplet (année, lemme, schéma morphosyntaxique, niveaux de dépendances) ;
– de filtres qui ne conservent que les formes choisies pour l'étude ;
– du noyau qui prend une liste de formes datées, propose un certain nombre de fonctions, et renvoie les résultats sous forme de texte à formater ou à utiliser dans une interface.

5.2. Les données traitées

5.2.1. Construction des données

Un corpus n'est pas une simple collection de textes, rassemblés au gré des disponibilités, mais un ensemble de textes, sélectionnés selon des critères linguistiques préétablis, avec un objectif de représentativité d'une partie de la langue (Sinclair, 1996). Le corpus d'étude a été construit à partir d'un ensemble de notices bibliographiques de physique extraites de la base *PASCAL* de l'INIST[2]. Ce corpus s'étend de 1984 à 1999. Il est composé d'environ un million de mots simples répartis dans les titres et les résumés de 6400 notices. Il peut être qualifié de *corpus de suivi spécialisé*, corpus de suivi parce qu'il est constitué de données textuelles qui s'accumulent au cours du temps (Habert *et al.*, 1998), corpus spécialisé par ce que ces données sont extraites d'articles scientifiques appartenant au même domaine et à la même communauté de discours.

Le corpus a été soumis aux étapes successives de nettoyage de texte, d'étiquetage grammatical (Brill, 1992), de lemmatisation avec *FLEMM* (Namer, 2000). Puis l'extraction des candidats-termes a été réalisée avec le logiciel *ACABIT* (Daille, 1994) en même temps qu'ont été calculés les niveaux de dépendance des termes et l'affectation des années.

5.2.2. Structuration dans le temps

Les résultats présentés ont été calculés avec le découpage temporel ci-contre, en choisissant la période n° 2 comme période de rupture. Le passé est donc composé des périodes 0, 1 et 2 et le futur des périodes 3 et 4.

Période	Intervalle	Nombre de termes
0 :	1984-1986	97
1 :	1987-1989	75
2 :	1990-1992	109
3 :	1993-1995	128
4 :	1996-1998	131

5.3. Résultats

5.3.1. Formes stables

Formes présentes dans chaque période du passé et dans chaque période du futur.

{diffusion de_le_hydrogène, diffusion de_neutron, diffusion diffus, diffusion du_atome, diffusion en_volume, diffusion incohérent, diffusion inélastique du_neutron, diffusion inélastique, diffusion multiple, diffusion neutronique, diffusion raman }

Formes dont au moins une variante est présente dans chaque période du passé et encore présente dans chaque période du futur.

[2] Nous remercions l'INIST, INstitut de l'information Scientifique et Technique situé à F-Vandœuvre les Nancy d'avoir mis ces données à notre disposition.

les formes de la liste précédente augmentées de : {diffusion atomique, diffusion cohérent, diffusion dans_le_volume, diffusion de_un_atome, diffusion du_neutron, diffusion hyper-raman, diffusion inélastique de_neutron, diffusion quasi élastique, diffusion volumique, diffusion élastique de_neutron, diffusion élastique du_neutron, diffusion élastique, diffusion de_atome}

5.3.2. Attestations nouvelles

Formes absentes de chaque période du passé et présentes dans chaque période du futur.

{diffusion cationique, diffusion de_ion, diffusion de_le_azote, diffusion diffus du_rayon, diffusion diffus observer, diffusion du_carbone, diffusion du_nickel, diffusion élastique de_le_lumière, diffusion lacunaire, diffusion quantique, diffusion simple, diffusion rayleigh}

Formes n'ayant aucune variante dans le passé et dont au moins une variantes est présente dans chaque période du futur.

{ diffusion al, diffusion de_agrégat, diffusion de_al, diffusion de_le_azote, diffusion de_un_agrégat, diffusion de_un_lacune, diffusion diffus observer, diffusion du_agrégat, diffusion du_carbone, diffusion du_nickel, diffusion lacunaire, diffusion quantique, diffusion quasi-élastique de_le_lumière, diffusion quasiélastique de_le_lumière, diffusion rayleigh, diffusion simple}

5.4. Analyse des résultats

5.4.1. Résultats sur les distances

Contrairement à beaucoup de travaux en TALN, il est ici impossible de comparer les résultats à des valeurs réputées justes. Le problème d'une expertise des distances se pose. Cependant ces distances n'ont pas de valeur intrinsèque et ne doivent être évaluées qu'au travers de leur rôle dans le contrôle des variantes. Un certain nombre de défauts reste à corriger :

– le problème de l'ordre des mots est incontournable sans un aménagement particulier ;
– la distance est sensible aux erreurs d'étiquetage grammatical ;
– les négations introduites par les préfixes, les prépositions ou les adverbes ne sont pas traitées ;
– la mise en place des niveaux de dépendance doit être remplacée par un système plus fin construit par exemple sur des arbres de dépendance ;
– les règles de calcul morphologique doivent être complétées ;
– les relations sémantiques ne sont pas traitées ;
– la distance est dépendante de la langue.

5.4.2. Résultats sur les suivis d'attestations terminologiques

L'ensemble des formes stables au sens strict est un sous-ensemble de celui des formes stables aux variantes près. La prise en compte des variantes permet de considérer par exemple que le terme *`diffusion du_neutron`* est stable même si sa forme exacte est absente de la période 1987-89. L'affichage du profil temporel de *`diffusion du_neutron`* permet de le vérifier.

1984-1986 : diffusion du_neutron, diffusion neutronique, diffusion de_neutron : 3
1987-1989 : diffusion neutronique, diffusion de_neutron : 2
1990-1992 : diffusion du_neutron, diffusion neutronique, diffusion de_neutron : 3
1993-1995 : diffusion du_neutron, diffusion neutronique, diffusion de_neutron : 3
1996-1998 : diffusion du_neutron, diffusion neutronique, diffusion de_neutron : 3

En ce qui concerne les nouvelles attestations, deux types de résultats doivent être examinés. Les calculs de suivis d'attestations ont été faits avec une rupture située après la période 1990-92.

Si un terme, comme *`diffusion cationique`*, est présent dans l'ensemble des apparitions

des formes exactes et pas dans celui des variantes, c'est qu'il présente au moins une variante dans le passé, donc ne peut être qualifié de nouveau. Son profil temporel permet de le confirmer.

1984-1986 : : 0	1993-1995 : diffusion cationique : 1
1987-1989 : : 0	1996-1998 : diffusion cationique, diffusion du_cation : 2
1990-1992 : diffusion du_cation : 1	

Si un terme, comme *diffusion al*, est présent dans l'ensemble des apparitions des variantes et pas dans celui des formes exactes c'est qu'il n'est pas apparu sous la même forme dans les périodes du futur. Son profil temporel le montre.

1984-1986 : : 0	1993-1995 : diffusion al : 1
1987-1989 : : 0	1996-1998 : diffusion de_al : 1
1990-1992 : : 0	

Une analyse plus générale des résultats sur le corpus d'étude montre que ceux-ci n'offrent pas des marques tangibles d'évolution. La raison est à chercher dans la constitution du corpus, tant il est encore difficile, à l'heure actuelle, de disposer de données électroniques suffisamment homogènes et s'étalant sur une durée significative. Cependant, la prise en compte de la variation terminologique permettra à coup sûr d'obtenir, sur des corpus de meilleure qualité, des conclusion plus fiables sur les suivis de termes.

6. Conclusion

Nous avons présenté un travail destiné à automatiser l'étude de l'évolution terminologique en exploitant des corpus diachroniques. La variation terminologique est prise en compte grâce à une distance entre termes complexes basée sur une adaptation de la distance d'édition entre chaînes. Les résultats obtenus apportent une information de meilleure qualité que si seules les formes exactes étaient considérées, puisqu'un terme peut être suivi dans le temps même s'il se présente sous des formes variantes. Deux applications peuvent alors être envisagées. L'une s'intéresse classiquement aux apparitions / disparitions d'attestations terminologiques, mais permet de compter avec les variantes. L'autre cherche à tracer l'histoire d'un terme particulier aux travers des différentes variantes sous lesquelles il apparaît. Ce travail attend des améliorations, en particulier pour affiner l'intégration des différents types de variations dans le calcul de distance.

Références

BENNETT C., GACS P. M. L., VITANYI M. et ZUREK W. (1998). « Information distance ». In *IEEE Transactions on Information Theory*, 44 (4), 1407–1423.

BRILL E. (1992). « A simple rule-based part of speech tagger ». In *Proceedings of the Third Conference on Applied Natural Language Processing*. p. 152-155.

CABRÉ M.-T. (1998). *La terminologie, théorie, méthodes et applications*. Armand Colin, Paris.

CALLON M., COURTIAL J.-P. et PENAN H. (1993). *La scientométrie*. Collection Que sais-je ? (2727). Presses Universitaires de France. épuisé.

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

DAILLE B. (2002). *Découvertes linguistiques en corpus*. Habilitation à diriger des recherches, Université de Nantes.

DE SAUSSURE F. (1916). *Cours de linguistique générale*. Payot, Paris. édité en 1969 par Charles Bally et Albert Sechehaye.

FAUVET M.-C. et SCHOLL P.-C. (1995). *Temps et Bases de Données. Concepts temporels pour la gestion de l'évolution des données*. rapport LGI, IMAG, Grenoble.

HABERT B., FABRE C. et ISAAC F. (1998). *De l'écrit au numérique. Constituer, normaliser et exploiter les corpus électroniques*. InterEditions.

JACQUEMIN C. (1999). « Syntagmatic and Paradigmatic representations of Term Variation ». In *Proceedings of 37th Annual Meeting of the Association for Computational Linguistics (ACL,99)*. University of Maryland, p. 341–348.

LEVENSHTEIN V. I. (1966). « Binary codes capables of correctiong deletions, insertions, and reversals ». In *Sov. Phys. Dokl.*, 10, 707–710.

MULLER C. (1992). *Principes et méthodes de statistique lexicale*. Champion-Slatkine, Paris-Genève.

NAMER F. (2000). « Un analyseur flexionnel du français à base de règles ». In *Traitement automatique des langues*, 41 (2), 523–548.

REY A. (1992). *La terminologie : noms et notions*. Collection Que sais-je ? (1780). Presses Universitaires de France, 2e edition.

SINCLAIR J. (1996). *Preliminary recommendations on Corpus Typology*. Rapport interne, EAGLES (Expert Advisory Group on Language Engineering Standards), CEE.

TARTIER A. (2004). *Analyse automatique de l'évolution terminologique : variations et distances*. Thèse de doctorat en informatique, Université de Nantes.

WAGNER R. et FISCHER M. (1974). « The string-to-string correction problem ». In *ACM*, 21, 168-173.