

Analyse automatique de l'évolution terminologique

Annie Tartier

Université de Nantes LINA FRE CNRS 2729
2 rue de la Houssinière
F-44322 NANTES Cedex 3
Annie.Tartier@univ-nantes.fr

Résumé

Cet article présente des méthodes de traitement automatique destinées à repérer des phénomènes d'évolution en analysant les termes extraits de corpus diachroniques spécialisés. Le premier axe consiste à repérer la nature des changements dans les attestations de certains termes. Cette analyse est effectuée en prenant en compte le phénomène de la variation terminologique, ceci grâce à la définition d'une distance entre formes terminologiques. Le deuxième axe concerne la structuration du temps et propose diverses modalités d'examen diachronique, destinées à distinguer les changements éphémères des changements durables qui, eux, pourraient être les signes d'une évolution. Les résultats obtenus sont tributaires du fait qu'il est encore difficile de construire des corpus diachroniques homogènes couvrant une période significative. Cependant les outils mis en place apportent les informations nécessaires au suivi de termes particuliers ou de cohortes de termes.

Mots-clés : Traitement automatique du langage naturel, terminologie, évolution, variation terminologique, distance.

Abstract

This paper presents methods for the automatic discovery of evolutionary phenomena within terms extracted from diachronic specialized corpora. The first axis consists in scanning which changes occur in the terms of the corpus. As a term has frequently variant forms, terminological variation is taken in account by the way of a distance defined between two terminological forms. The second axis concerns time structuration and proposes several diachronic examination modes in order to distinguish ephemeral changes from durable ones, which could be the signs of an evolution. Results are strongly dependent on the fact that it is still difficult to build homogenous diachronic corpora along a significant time period. But the computing system gives the information necessary to follow a term or a set of terms.

Key-words: Natural language processing, terminology, evolution, terminological variation, distance.

1 INTRODUCTION

Définissons le cadre de cet article par deux citations d'Alain Rey [23].

« *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. »

Différentes catégories de professionnels sont concernées par l'évolution des terminologies. Les premiers, scientifiques et techniciens, sont directement impliqués dans la création des termes, ainsi que les traducteurs et rédacteurs qui en assurent la diffusion. Viennent ensuite les professionnels de la terminologie, qui prennent des décisions techniques et politiques au sujet des termes à utiliser. Il y a, enfin, ceux qui font des termes leur objet d'étude. Parmi ces derniers les acteurs de la veille et les épistémologues cherchent des informations sur l'évolution des domaines au travers de l'évolution des termes, les uns dans le cadre d'une production industrielle, les autres avec des objectifs de connaissance. Dans cette dernière catégorie on trouve aussi les linguistes dont l'intérêt porte sur le fonctionnement de la langue.

Malgré des objectifs différents, les besoins se rejoignent en terme d'analyse de la production langagière ou terminologique. Notre propos 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 donc de mettre au jour les évolutions dans les textes et, en particulier, de suivre au cours du temps ce que devient un terme, et quels sont les termes stables ou instables.

Après un bref panorama des travaux dans ce domaine, nous analysons les deux volets du problème, d'une part la nature de ce qui change, d'autre part la distribution temporelle des changements. En nous appuyant sur ce que recouvrent les termes *changement* et *évolution*, nous expliquons comment il est possible d'organiser une analyse diachronique automatique, puis nous montrons que celle-ci est insuffisante en l'absence de prise en compte du phénomène de la variation terminologique. C'est par le biais d'une « distance » entre formes terminologiques que celle-ci est intégrée à l'analyse diachronique. Les méthodes présentées sont implémentées dans un prototype et quelques résultats, obtenus à partir d'un corpus diachronique, sont donnés.

2 ÉTUDES DIACHRONIQUES EN TERMINOLOGIE

2.1 Une « terminologie diachronique »

N'y a-t-il pas, à l'instar de la linguistique diachronique, dissociée de la linguistique synchronique par F. de Saussure [10], une vue diachronique de la terminologie. Le terme « terminochronie » a d'ailleurs été proposé par Møller [19] en 1998. Les travaux associés à ce domaine sont disparates. Les uns analysent et décrivent des points particuliers de l'évolution d'un terme. D'autres, plus systématiquement, construisent des outils ou des ressources. Cette tendance est favorisée par la disponibilité de plate-formes informatiques et de grandes capacités de stockage et de structuration de données.

2.2 La néologie

Les terminologues n'ont pas attendu les outils informatiques pour recenser et caractériser les nouveaux termes. Les outils de la néologie proposent traditionnellement deux types de fonctionnalités : le recensement et la mémorisation, réalisés par des spécialistes en terminologie, ainsi que la diffusion et l'interrogation qui permettent aux usagers de terminologies d'utiliser les connaissances mémorisées. S'ajoutent à ceux-ci, et de plus en plus, les outils de détection automatique à partir de corpus de textes. À titre d'illustration nous citons quelques outils :

Le système Balnéo [16] créé en 1994 par le *RINT*¹, en collaboration avec le laboratoire *CRAIE*² de l'Université de Rennes II, a pour objectifs la collecte, l'échange et la diffusion rapides de matériaux terminologiques touchant plus particulièrement les néologismes, afin de rendre plus facile et plus efficace la mise à jour des dictionnaires terminologiques et les banques de terminologie.

BORNÉO est une base d'observation et de recherche des néologismes mise en place par l'ex *INaLF*³ (actuellement *ATILF*⁴) qui présente les unités néologiques en contexte de manière chronologique. Ces unités sont relevées dans des énoncés de presse contemporaine et sont considérées comme « néologiques » dès lors qu'elles ne figurent pas dans un dictionnaire d'usage. Cette base permet d'analyser les ressources et la créativité dénomminative du français du dernier quart du 20^e siècle.

CENIT (Corpus-based English Neologism Identifier Tool) [24] est un système de détection semi-automatique de néologismes. Ce prototype filtre tous les mots d'un texte d'entrée au travers d'un corpus d'exclusion.

La liste n'est pas close. Des outils comme *Webaffix* [26] puisent dans les ressources du Web pour acquérir des unités lexicales absentes de lexiques de référence et donc supposées nouvelles.

¹Réseau International de Néologie et de Terminologie.

²Centre de Recherche et d'Application en Ingénierie linguistiqueE.

³Institut National de la Langue Française

⁴Analyse et Traitement Informatique de la Langue Française

2.3 Corpus diachroniques et évolution du lexique

2.3.1 Corpus diachroniques

L'émergence de la linguistique de corpus et la disponibilité d'outils automatiques d'exploitation ont changé la dimension de la linguistique diachronique. Le chapitre « Le langage au fil du temps : corpus et diachronie » de Habert et al. [14] fait le point sur la structure et les usages de corpus diachroniques. Il met en garde sur le fait qu'un corpus structuré par l'écoulement du temps n'est pas forcément un *bon* corpus diachronique. Lebart et Salem [17] utilisent le terme de *séries textuelles chronologiques* pour désigner des corpus diachroniques complètement homogènes par tous les caractères autres que l'époque. La disponibilité de telles ressources est une aide incontestable pour l'étude de l'évolution du vocabulaire. Des corpus diachroniques sont construits pour tenter d'établir avec certitude des faits d'évolution de la langue. Ceux de langue anglaise (Helsinki, Archer, Cambridge, ...) sont les plus nombreux. Les quelques corpus de langue française présentés ci-dessous sont choisis volontairement dans différents domaines, pour comparer les types de construction et les méthodes d'exploitation diachronique.

2.3.2 Un corpus du français littéraire

La base *FRANTEXT*, construite par l'*INaLF* contient environ 3500 textes littéraires, qui s'échelonnent du 16^e au 20^e siècle. Avec plus de 150 millions de mots *FRANTEXT* est une source lexicale extrêmement riche pour mener des études diachroniques du français littéraire. Un logiciel d'exploitation, *THIEF*⁵, lui est associé. Conçu par Étienne Brunet pour offrir des outils de traitement quantitatif de la base *FRANTEXT*, certaines fonctionnalités ont délibérément un objectif diachronique. Une base locale comprend le relevé de toutes les formes et de leurs fréquences dans douze tranches chronologiques calculées pour être de tailles à peu près semblables. Les formes brutes des mots simples sont soumises à des traitements de statistique lexicale [21].

2.3.3 Un corpus structuré par des événements historiques

Belica [3] a travaillé sur un corpus d'articles de journaux et de discours dont la particularité est de s'étendre sur une très courte période (du 6 mai 1989 au 31 décembre 1990). On pourrait douter de la possibilité d'observer une évolution du lexique sur une période aussi courte, si ce n'est que cette période est fortement marquée par l'événement important qu'est la chute du mur de Berlin (novembre 1989). Les quantités de textes venant de l'Ouest (ex RFA) et de l'Est (ex RDA) sont équivalentes. Il est découpé en sept tranches chronologiques dont les limites sont déterminées par des événements importants de la période. L'objectif est de mettre au jour des irrégularités diachro-

⁵Truchement Hypertexte pour l'Interrogation et l'Exploitation de Frantext.

niques en calculant s'il existe des différences notables dans la répartition de certaines chaînes dans les tranches.

2.3.4 Un corpus de discours politiques

Le travail rapporté par Monnière [20] concerne un corpus constitué de tous les discours inauguraux⁶ prononcés chaque année, de 1945 à 1996, devant l'Assemblée du Québec. Il comprend 49 discours, avec un total de 191724 occurrences de mots. Pour neutraliser le fait que les discours sont de longueurs très variables, le corpus a été divisé en portions de 1000 occurrences de mots. Les mesures faites sur ce corpus portent sur trois points : l'accroissement lexical d'une tranche à l'autre, la distance lexicale entre deux textes et des caractérisations lexicales, jugées en comparant la fréquence d'un mot dans une partie à la fréquence dans l'ensemble du corpus.

2.3.5 Comparaison de deux corpus distants dans le temps

Il s'agit d'un travail, présenté par Aussenac [2], ciblé sur l'évolution terminologique du domaine de l'ingénierie des connaissances. L'objectif est de repérer des évolutions thématiques au travers de l'évolution des termes. La méthode suivie est celle d'une analyse contrastive de deux corpus datant de deux périodes différentes (1995-1998 et 1999-2001), orientée selon deux axes : les contenus et les contextes. Les candidats termes sont extraits à l'aide du logiciel *SYNTEX* qui fournit leur fréquence et leur répartition. L'évolution des contenus est étudiée en terme de hausse ou de baisse à partir de comparaisons de ces valeurs. Les contextes sont étudiés sous la forme de coefficients de proximité distributionnelle calculés par le logiciel *UPERY*.

2.4 Synthèse

Malgré le peu d'exemples, il apparaît que les méthodes d'exploitation des corpus passent toutes par une stratégie de segmentation en tranches chronologiques, guidée par la nature du corpus et par l'objectif de l'étude. Les corpus sont ensuite réduits en unités atomiques, formes brutes des mots ou formes lexicales complexes. Des indices d'évolution sont alors calculés à partir de comptage d'occurrences ou de co-occurrences de ces unités de base.

Peut-on encore parler de corpus diachronique lorsque la segmentation est calée sur des événements ? Ceux-ci, en effet, ne sont pas liés au temps absolu et auraient pu avoir lieu à d'autres dates. Ceci met en évidence le fait que le temps n'est qu'un support des événements qui sont, eux, responsables de l'évolution. Or le temps est utilisé comme un indicateur parmi d'autres, ce qui pose un problème de fond parce que, s'il est lié aux données, il n'en est pas forcément un bon représentant.

⁶Tirés des Journaux de l'Assemblée législative de la province de Québec de 1944 à 1971 et des Procès-verbaux de l'Assemblée nationale du Québec de 1971 à 1996.

3 UN PROBLÈME À DOUBLE ENTRÉE

Lorsqu'il est question d'exprimer l'évolution d'un système, l'idée première est de projeter des éléments caractéristiques de ce système sur l'axe du temps. Une partie de l'analyse concerne donc l'espace du système et doit déterminer ce qui est modifiable dans la nature de ses éléments. Une autre partie décide de la manière de projeter sur le temps pour que ces modifications prennent une signification par rapport à la question de l'évolution. La première partie, qui consiste à analyser des objets tangibles, peut être conduite indépendamment de la seconde (bien que la seconde soit implicitement sous-jacente), alors qu'il est difficile de traiter de projection sur le temps si les entités à projeter n'ont pas été définies.

3.1 Définition des objets observés

Dans les définitions proposées par la littérature [23, 7, 11] pour ne citer qu'eux, le *terme* est un objet à deux facettes : une forme linguistique et une fonction de référence à une notion ou à un concept. Seules les formes linguistiques des termes⁷, peuvent être analysées directement par un programme, l'interprétation sur les concepts restant le fait des linguistes ou des terminologues. Un ensemble de formes terminologiques, dont chacune est marquée par une date d'attestation (date à laquelle elle a été trouvée dans un texte), constitue le matériau d'observation au travers duquel sont recherchées des marques d'évolution. Deux niveaux d'étude se dessinent : au niveau microscopique c'est l'histoire d'un terme que l'on cherche à tracer, au niveau macroscopique l'objectif est de suivre des cohortes de termes, regroupés sur des critères préétablis comme l'appartenance à un domaine, à une communauté de discours ou, ayant tout autre caractéristique commune susceptible de porter une signification.

3.2 Changement

À une date donnée, la manifestation la plus simple du changement est l'apparition ou la disparition⁸ d'une forme terminologique. La situation serait simple si la disparition d'une forme pouvait induire la disparition de l'usage du terme référent. Or il n'en est rien puisque'une forme d'un terme peut disparaître alors qu'une forme variante du même terme continue d'être attestée. Il importe donc d'être capable de reconnaître un terme sous ses formes variantes avant de statuer sur son devenir. Pour cette raison il faut disposer d'un outil capable de décider si deux formes terminologiques données sont les variantes d'un même terme. C'est dans ce but qu'a été définie une « distance morphosyntaxique » entre deux formes terminologiques (voir section 4).

⁷Appelées plus simplement *formes terminologiques*.

⁸Par abus de langage, apparition (resp. disparition) d'une forme terminologique signifie apparition (resp. disparition) de son attestation.

3.3 Évolution

Changement et évolution ne sont pas synonymes. Certains changements sont éphémères. Il n'y a évolution que lorsque le changement est confirmé. Exprimer l'évolution ne consiste donc pas uniquement à repérer des occurrences de changements. Pour repérer un changement il faut observer et prendre des mesures sur les choses alors que pour repérer une évolution il faut caractériser les changements au cours du temps. Cette idée fonde les modalités d'observation temporelle exposées à la section 5.

3.4 Démarche

Le premier volet de notre démarche (figure 1) est une analyse du phénomène de la variation terminologique sur laquelle s'appuie la définition d'une « distance morphosyntaxique » entre deux formes terminologiques (section 4). Disposer d'une telle mesure permet de regrouper les variantes d'un terme, aussi bien au niveau microscopique (suivi d'un terme) que macroscopique (suivi d'une cohorte de termes). Le deuxième volet, présenté à la section 5, propose des modalités d'observation diachronique conçues pour ne pas réduire l'analyse de l'évolution à la détection d'occurrences de changements.

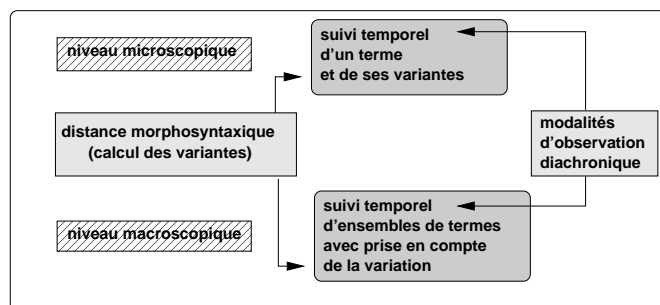


FIG. 1 – Démarche

4 CALCUL DES VARIANTES

Décider, indépendamment de tout contexte, que deux formes terminologiques sont probablement les variantes d'un même terme est une question cruciale que nous cherchons à résoudre en définissant une *mesure* destinée à quantifier la *variation* entre les deux formes. Il devient alors possible de qualifier automatiquement de *variantes* deux formes pour lesquelles cette mesure est inférieure à un seuil, paramétrable selon l'application envisagée. Il n'est bien sûr pas question de réduire un phénomène linguistique à un nombre, mais la disponibilité d'une telle mesure rend opératoire la détection des variantes indépendamment de tout contexte.

4.1 Variation terminologique : une typologie

Il existe de nombreuses typologies de la variation [15, 9], conçues le plus souvent en fonction des objectifs de leurs auteurs. La méthode de calcul de distance présentée ici s'appuie sur la typologie suivante, extraite de [9] :

- variation orthographique (*quasiélastique / quasi-élastique*)
- variation morphologique : dérivation, composition (*photoélectron / électron*)
- variation syntaxique : syntaxique faible (*diffusion laser / diffusion par laser*), ellipse d'un nom support (*diffraction électronique rasante / diffraction électronique en incidence rasante*), insertion, expansion, et substitution (*diffraction de neutrons / diffraction élastique de neutrons*)
- variation morphosyntaxique faisant intervenir des dérivations morphologiques (*diffraction électronique / diffraction des électrons*)

Le traitement des abréviations fait partie des projets d'enrichissement de cette liste. En revanche, les « variations sémantiques » (synonymes, hyperonymes), ne peuvent pas être traitées uniquement à partir des formes et ne sont pas, pour l'instant, intégrées à ce travail.

4.2 « Distance » entre deux formes terminologiques

La *distance informationnelle* [4] entre deux entités complexes est définie par la longueur du plus court programme permettant de décrire la transformation d'une entité dans l'autre. Cette définition 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 [18], qui est justement fondée sur le principe de la transformation, puis nous l'avons adaptée à la comparaison de deux termes complexes.

4.3 Distance d'édition entre chaînes

Le mécanisme de base est une transformation d'une chaîne en une autre par application d'opérations élémentaires sur certains caractères (insertion, suppression, substitution). Un coût est attribué à chaque opération. Chaque transformation globale, résultant des opérations élémentaires, se traduit par un alignement des deux chaînes. Le coût d'un alignement est égal à la somme des coûts des opérations élémentaires et le coût minimal constitue la distance recherchée. Il importe 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 [28].

4.4 Adaptation aux termes complexes

4.4.1 Principe

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

<i>diffusion cohérente inélastique d'électrons</i>
<i>diffusion cohérente</i> des neutrons thermiques
1 suppression 1 substitution 1 insertion

Mais cette adaptation ne peut être une simple transposition. Pour la réaliser, 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 terme⁹,
- 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 .

terme	:	<i>diffusion inélastique de neutrons thermiques</i>			
lemme	:	diffusion	inélastique	de_neutron	thermique
schéma	:	SBC	ADJ	PREP_SBC	ADJ
niveaux	:	0	1	1	2

Le principe de l'algorithme est le même que celui de la distance d'édition, mais les coûts élémentaires sont calculés de manière à ce que les variations induisent des alignements pertinents et influent sur la distance finale.

4.4.2 Démarche suivie pour le calcul des coûts élémentaires

Une échelle de *coûts de base* est établie pour chaque type d'opération et, dans le cas des substitutions, pour chaque type de variation. Comme les choix de valeurs s'appuient nécessairement sur une « interprétation linguistique », les valeurs de ces coûts sont paramétrables. Aux extrémités de l'échelle se trouvent un coût nul pour la substitution de deux constituants identiques et un coût très élevé pour interdire certains alignements en privilégiant une insertion et une suppression au détriment d'une substitution. Les autres coûts de substitution, classés du plus faible au plus fort, sont guidés par la typologie de variation présentée en 4.1.

1. éléments qui se correspondent par une variation syntaxique faible (*de_neutron / du_neutron*),
2. éléments qui se correspondent par une variation morphosyntaxique (*neutronique / du_neutron*),

⁹Étiquettes de Brill [6] : SBC (substantif commun), ADJ (adjectif), PREP (préposition), ...

3. éléments différents présentant une similitude dans le schéma syntaxique (*de_neutron / de_électron*).
4. éléments ne présentant aucune similitude (*de_neutron / électronique* ou *de_neutron / par_laser*).

Les coûts d'insertion et de suppression sont positionnés au niveau 3 de cette échelle. Les liens morphologiques sont repérés par des calculs à base de règles et des listes d'affixes. Dans la substitution de deux éléments tels que *de photo-électron* et *du photoélectron*, il apparaît simultanément une variation syntaxique faible (*de / du*) et une variation orthographique (trait d'union). Lorsqu'il y a superposition de variations sont indépendantes, les coûts correspondants sont ajoutés

Les niveaux de dépendance servent à pondérer les opérations qui s'appliquent aux constituants qui ne sont pas en position de tête. Par ailleurs, pour obtenir un alignement cohérent, il faut forcer la substitution des deux éléments de tête tout en donnant une grande valeur au coût de substitution si les têtes sont différentes. Pour qu'un tel coût de substitution n'écarte pas la substitution au profit d'une insertion et d'une suppression, les opérations portant sur les éléments de tête sont affectées d'une valeur de pénalité, paramétrable comme les autres valeurs de coûts.

Au cours de l'exécution de l'algorithme de programmation dynamique, ce sont les étiquettes syntaxiques des constituants qui aiguillent sur les règles de calcul appropriées. À titre d'exemple, la substitution de deux éléments d'étiquettes *PREP_SBC* conduit à examiner des cas comme :

$SBC_1 = SBC_2$ et $PREP_1 = PREP_2$	formes identiques
$SBC_1 = SBC_2$ et $PREP_1 \neq PREP_2$	variation syntaxique faible
$SBC_1 \approx SBC_2$ et $PREP_1 = PREP_2$	variation morphologique
$SBC_1 \approx SBC_2$ et $PREP_1 \neq PREP_2$	variation syntaxique faible + morphologique
...	

4.4.3 Exemples

Les « distances » de l'exemple ci-dessous ne sont pas normalisées et ne jouent pas le rôle de « distances sémantiques » entre les termes. Elles ne servent qu'à rendre opérationnelle la reconnaissance de variantes par comparaison avec un seuil paramétrable (placé par exemple au niveau du trait).

<i>diffusion de_neutron, diffusion neutronique</i>	0.50
<i>diffusion neutronique, diffusion élastique de_neutron</i>	1.50
<i>diffusion lent de_neutron, diffusion du_neutron lent</i>	1.75
<i>diffusion de_neutron, diffraction de_neutron</i>	7.50

5 PROJECTION SUR LE TEMPS

La segmentation d'un corpus par le temps agit comme un outil d'analyse. Elle ne devrait pas induire d'information extérieure aux données.

5.1 Un modèle à temps discret

Le temps n'est qu'un repère qui permet de définir une relation d'ordre total (chronologique) entre des événements. Il existe différents modèles de représentation du temps [1, 5]. La nature des données (formes terminologiques marquées par l'année du texte dans lequel elles apparaissent) conduit à choisir un modèle à temps discret, dans lequel l'axe des temps est une suite d'intervalles unitaires consécutifs, les *chronons* [12], munis d'une relation d'ordre. Un intervalle quelconque est la réunion d'intervalles élémentaires consécutifs. Une bijection entre un sous-ensemble d'intervalles consécutifs et un sous-ensemble d'entiers consécutifs permet d'associer une *date* à chaque intervalle élémentaire de sorte qu'à deux intervalles consécutifs, correspondent deux entiers consécutifs.

Il est donc équivalent de dire qu'un prédicat est vrai sur un intervalle élémentaire ou à une certaine date. La fusion de plusieurs intervalles élémentaires consécutifs I_i, I_{i+1}, \dots, I_j associés aux dates entières consécutives d_i, d_{i+1}, \dots, d_j produit un intervalle non élémentaire $[d_i, d_j]$ qui peut être défini soit par les deux dates d_i et d_j associées au premier et au dernier intervalle élémentaire, soit par un couple (d_i, duree) où $\text{duree} = d_j - d_i + 1$ est un entier égal au nombre d'intervalles élémentaires fusionnés.

5.2 Faits, événements

Un *fait*, avéré ou non à une certaine date, est un prédicat dont la valeur de vérité dépend de la date. Le *fait* typique est l'attestation ou la non attestation d'une forme terminologique sur un certain intervalle, élémentaire ou non.

Pour garder la cohérence de la représentation du temps, un *événement* n'a pas de durée, c'est une action qui a lieu ponctuellement entre deux intervalles élémentaires comme par exemple la *disparition* de la forme f entre d_2 et $d_2 + 1$ si f est attestée sur $[d_1, d_2]$ et non attestée sur $[d_2 + 1, d_3]$. Les événements fondamentaux sont de trois types : le maintien d'une forme f qui est une sorte d'« événement neutre », la disparition et l'apparition d'une forme.

Rappelons que notre rôle se limite à l'observation d'événements sur les attestations dans un corpus donné. Nous ne concluons pas sur la nouveauté ou la perte d'usage d'un terme, même si, pour des raisons de concision, les mots utilisés comme *disparition* ou *apparition*, pourraient le laisser croire.

5.3 Périodes

Les données sont constituées d'un ensemble de formes terminologiques datées partitionné en *périodes* munies d'un *ordre chronologique*. Une *période* est un ensemble de formes dont les dates d'attestations appartiennent à un intervalle (*date_début*, *durée*) nommé *en-tête temporel*. L'attestation d'une forme terminologique à une certaine date est modélisée par son appartenance à une période « couvrant » cette date. Chaque période est affectée

d'un indice entier tel que l'ordre numérique des indices corresponde à l'ordre chronologique des périodes, et qu'à deux indices consécutifs dans les entiers correspondent deux périodes consécutives dans le temps. La réunion de toutes les périodes consécutives, c'est-à-dire l'ensemble de toutes les formes terminologiques étudiées, est nommée l'*époque*.

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.

5.4 Passé, futur, rupture

Pour repérer un événement il faut comparer un état d'*avant* et un état d'*après* cet événement. Si la vue temporelle est composée de n périodes de \mathcal{P}_1 à \mathcal{P}_n , l'une d'entre elle \mathcal{P}_t est désignée comme *période de rupture*, définissant ainsi le *passé* comme la réunion des périodes consécutives allant de \mathcal{P}_1 à \mathcal{P}_t et le *futur* comme la réunion des périodes consécutives allant de \mathcal{P}_{t+1} à \mathcal{P}_n . Étudier l'évolution de l'ensemble des termes revient à mettre en évidence des *modifications* ou des *conservations* avant et après la rupture. L'introduction d'une rupture est incontournable, mais elle doit être mobile et permettre l'observation de toutes les configurations (passé, futur) possibles. Enfin, la fusion de certaines périodes consécutives en périodes plus larges doit être permise. Une réponse à la question de la durabilité des faits posée au paragraphe 3.3 peut maintenant être envisagée.

5.5 Modalités d'examen diachronique

Il ne suffit pas de comparer des occurrences de faits dans le passé et dans le futur, il faut examiner leur distribution temporelle pour rendre compte de leur caractère occasionnel ou permanent.

L'époque est déterminée par l'intervalle $[EpDeb, EpFin]$ où $EpDeb$ et $EpFin$ désignent l'indice de la première et de la dernière période. Par convention l'indice de rupture t est celui de la dernière période du passé. La rupture peut prendre une position quelconque, $\forall t \in [EpDeb, EpFin]$.

On établit plusieurs modes d'examen diachronique. Tous ne considèrent pas la totalité de l'époque, mais une *vue temporelle* particulière qui doit contenir au moins la période de rupture et sa suivante. Les indices de sa première et de sa dernière période sont notés $Vdeb$ et $Vfin$. La structure de cette vue temporelle définit les modes d'examen diachronique. La figure 2 montre les quatre modes qui ont été implémentés.

Dans le *mode local (L)* où $[Vdeb, Vfin] = [t, t + 1]$, on compare la réalisation d'un fait dans deux périodes consécutives.

Dans le *mode consolidé dans le futur (CF)* où $[Vdeb, Vfin] = [t, EpFin]$, la question est de savoir si un fait d'une certaine période \mathcal{P}_t se confirme ou s'infirme de manière non occasionnelle dans le futur de celle-ci, c'est à dire dans *chaque* période $\mathcal{P}_k, \forall k \in [t + 1, EpFin]$.

Le *mode consolidé dans le passé (CP)* est symétrique du précédent. On veut savoir si un fait d'une certaine période \mathcal{P}_{t+1} existait déjà de manière non occasionnelle dans le passé de celle-ci, c'est à dire dans *chaque* période $\mathcal{P}_i, \forall i \in [EpDeb, t]$.

Le *mode permanent (P)* rassemble les deux précédents. Un fait non occasionnel du passé se confirme-t-il ou s'infirmes-t-il de manière non occasionnelle dans le futur. Il faut donc examiner si le fait se réalise dans *chaque* période \mathcal{P}_i du passé, $\forall i \in [EpDeb, t]$ et dans *chaque* période \mathcal{P}_k du futur, $\forall k \in [t+1, EpFin]$.

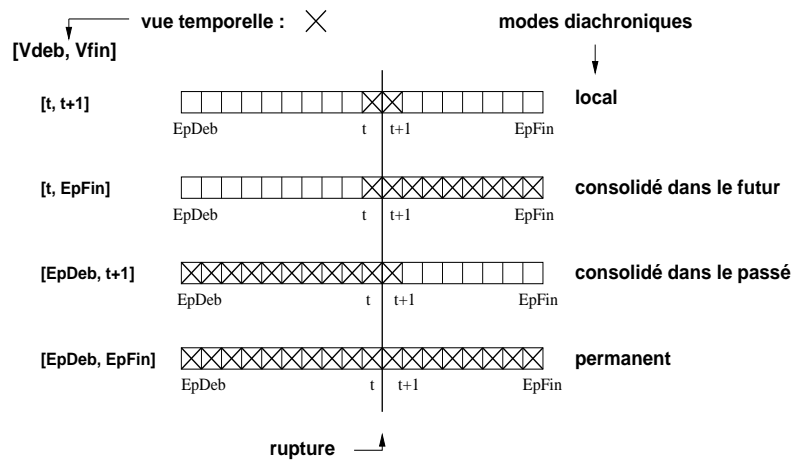


FIG. 2 – Modalités d'examen diachronique

Ces modalités sont rigides et doivent être affinées. Il est toujours possible de travailler sur une vue quelconque pourvu que la rupture et la période suivante lui appartienne. Il faut alors disposer de critères permettant de définir cette vue car il n'est pas envisageable de générer toutes les vues possibles et de ne conserver que celles qui sont fructueuses. Pratiquement c'est la possibilité d'élargir ou de fusionner certaines périodes qui permet de pallier les contraintes trop fortes imposées par les modalités diachroniques proposées.

6 EXTENSION AUX VARIANTES

6.1 Périodes étendues aux variantes

Sur le plan mathématique les périodes sont des ensembles. En plus des trois opérations classiques d'intersection, de réunion et de différence, une opération, spécifique au problème, 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, $\mathcal{V}_{\mathcal{P}_i}^{\mathcal{P}_k}$, constitue l'ensemble des variantes de \mathcal{P}_i

dans \mathcal{P}_k (1). La fonction $variantes(f_k, f_i)$ compare à un seuil la distance entre f_k et f_i pour décider si celles-ci sont des variantes. $\mathcal{P}_i \cup \mathcal{V}_{\mathcal{P}_i}^{\mathcal{P}_k}$ est la période \mathcal{P}_i étendue à ses variantes dans \mathcal{P}_k .

$$\mathcal{V}_{\mathcal{P}_i}^{\mathcal{P}_k} = \{f_k \in \mathcal{P}_k \setminus \mathcal{P}_i, \exists f_i \in \mathcal{P}_i \text{ et } variantes(f_k, f_i)\} \quad (1)$$

Une illustration est donnée à la figure 3 considérant des variantes morpho-syntaxiques (adjectif-nom) et syntaxiques faibles.

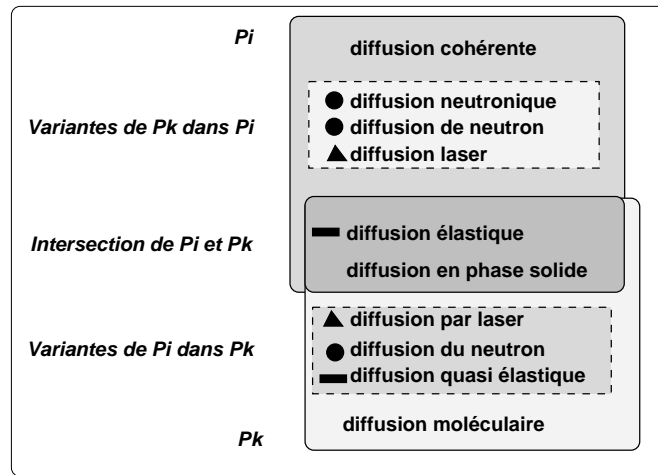


FIG. 3 – Périodes étendues aux variantes

6.2 Modalités d'examen des événements

Les événements fondamentaux sont la *stabilité*, la *disparition*, et l'*apparition* de l'attestation de formes terminologiques. Chacun de ces trois événements est envisagé selon deux modalités notées *S* comme *au sens strict* et *V* comme *aux variantes près*. Dans la modalité *S* l'événement porte sur la forme exacte du terme, dans la modalité *V*, l'événement porte sur la forme exacte ou l'une de ses variantes. Ce deuxième mode considère comme équivalentes :

- la présence d'un terme ou d'une de ses formes variantes,
- l'absence d'un terme et de toutes ses formes variantes.

Il permet, par exemple, de ne pas déclarer qu'un terme est devenu obsolète alors qu'il continue d'apparaître sous une forme variante.

6.3 Expression des événements

6.3.1 Considérations générales

Les résultats sont des ensembles résultant d'intersections, d'unions et de différences de périodes ou de périodes étendues aux variantes. Une *intersection généralisée* entre des périodes renvoie l'ensemble de toutes les formes présentes dans chacune des périodes (stabilité). Une *union généralisée* renvoie l'ensemble des formes présentes dans au moins une période. La *différence* entre une intersection de périodes \mathcal{P}_i et une union de périodes \mathcal{P}_k renvoie l'ensemble de toutes les formes qui appartiennent à chaque période \mathcal{P}_i et à aucune période \mathcal{P}_k (apparitions et disparitions). Les définitions qui suivent s'entendent toutes « entre le passé et le futur » sous l'un des modes d'examen diachronique, déterminé par le choix de la *vue temporelle*, et indépendamment de la position de la rupture. Les expressions des ensembles « au sens V » sont obtenues en remplaçant les périodes \mathcal{P} par les périodes étendues $\mathcal{P} \cup \mathcal{V}_P^{Vue}$ dans les expressions des ensembles « au sens S ».

6.3.2 Stabilité de l'attestation

Une forme est stable *au sens strict* si elle est attestée dans chaque période du passé et dans chaque période du futur, c'est-à-dire dans chaque période de la vue temporelle. Un terme est stable *aux variantes près* si au moins une de ses variantes est présente dans tout le passé et dans tout le futur. Les ensembles S_e (au sens strict) S_v (aux variantes près) sont donnés en (2).

$$S_e = \bigcap_{i=Vdeb}^{Vfin} \mathcal{P}_i \quad S_v = \bigcap_{i=Vdeb}^{Vfin} \left(\mathcal{P}_i \cup \mathcal{V}_{\mathcal{P}_i}^{Vue} \right) \quad (2)$$

6.3.3 Suivi des variantes

Pour affiner l'observation de la cohorte des formes stables, il est intéressant de mettre en évidence des variations qui ont lieu au passage de la rupture. On nomme *variation diachronique* tout couple (f_x, f_y) de variantes dont le premier élément est une forme présente dans chaque période du passé, et le second dans chaque période du futur. L'expression (3) renvoie les *variations diachroniques*.

$$\left\{ (f_x, f_y) \mid \text{variantes}(f_x, f_y) \text{ et } (f_x, f_y) \in \left(\bigcap_{i=Vdeb}^t \mathcal{P}_i \right) \times \left(\bigcap_{k=t+1}^{Vfin} \mathcal{P}_k \right) \right\} \quad (3)$$

6.3.4 Disparition d'attestations

Une forme terminologique a disparu *au sens strict* si elle est attestée dans chaque période du passé et non attestée dans chaque période du futur. L'en-

semble D_e des ces formes est donné en (4).

$$D_e = \left(\bigcap_{i=Vdeb}^t \mathcal{P}_i \right) \setminus \left(\bigcup_{k=t+1}^{Vfin} \mathcal{P}_k \right) \quad (4)$$

Dans ce cas il se peut que la forme exacte ne soit plus attestée, mais que certaines de ses variantes persistent dans le futur. Pour décider qu'un terme est obsolète, il faut qu'une de ses variantes soit présente dans chaque période du passé et qu'aucune de ses formes variantes ne soit présente dans une quelconque période du futur. Ces ensembles D_v sont présentés en (5).

$$D_v = \left(\bigcap_{i=Vdeb}^t (\mathcal{P}_i \cup \mathcal{V}_{\mathcal{P}_i}^{Vue}) \right) \setminus \left(\bigcup_{k=t+1}^{Vfin} (\mathcal{P}_k \cup \mathcal{V}_{\mathcal{P}_k}^{Vue}) \right) \quad (5)$$

6.3.5 Nouvelles attestations

Des formules analogues dans lesquelles le rôle du passé et du futur est échangé, donnent les nouvelles attestations.

7 RÉSULTATS

7.1 Le prototype

Les résultats présentés dans cette section ont été calculés grâce aux fonctions d'un prototype destiné à analyser l'évolution des terminologies [27]. Celui-ci est composé de modules de pré-traitement, dédiés à chaque source de données, de filtres qui ne conservent que les formes choisies pour l'étude, et du noyau qui prend une liste de formes datées et renvoie les résultats sous forme de texte à formater ou à utiliser dans une interface.

7.2 Les données traitées

Les textes composant un corpus ne doivent pas être simplement rassemblés au gré des disponibilités, mais sélectionnés selon des critères linguistiques préétablis, avec un objectif de représentativité d'une partie de la langue [25]. 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*¹⁰. 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* parce qu'il est constitué de données textuelles qui s'accumulent au cours du temps [13], et que ces données sont extraites d'articles scientifiques appartenant au même domaine.

¹⁰Nous remercions l'INIST, INstitut de l'information Scientifique et Technique situé à Vandœuvre les Nancy d'avoir mis ces données à notre disposition.

Le corpus a été soumis aux étapes successives de nettoyage de texte, d'éti-quetage grammatical [6], de lemmatisation [22]. Puis l'extraction des candidats-termes a été réalisée avec le logiciel *ACABIT* [8] en même temps qu'ont été calculés les niveaux de dépendance des termes et l'affectation des années.

7.3 Un extrait des résultats

Au niveau macroscopique les résultats se présentent sous forme de listes. Un extrait est donné avec la période n° 2 comme période de rupture.

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

7.3.1 Formes stables

Au sens strict : 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 }

Aux variantes près : formes dont au moins une variante est présente dans chaque période du passé et dans chaque période du futur.

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 }

7.3.2 Attestations nouvelles

Au sens strict : 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 }

Aux variantes près : formes n'ayant aucune variante dans le passé et dont au moins une variante 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 }

7.4 Analyse des résultats

Contrairement à beaucoup de travaux en TALN, il est ici impossible de comparer les résultats des distances 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.

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

Dans l'observation des nouvelles attestations, 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.

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 conclusions plus fiables sur les suivis de termes.

8 CONCLUSION

Nous avons présenté des méthodes destinées à automatiser l'étude de l'évolution terminologique en exploitant des corpus diachroniques. Elles complètent les travaux de ce type principalement sur deux points. Grâce à la définition d'une « distance morphosyntaxique » entre termes complexes, nous

limitons le bruit occasionné par le phénomène de la variation terminologique. Ceci permet d'avoir une vue plus juste de l'évolution réelle de l'attestation terminologique. La mise en place d'une stratégie de segmentation par le temps, paramétrable, permet de distinguer l'occasionnel du permanent, ce qui est primordial dès qu'il s'agit de tracer un terme ou un ensemble de termes. Ces méthodes doivent être affinées en même temps que doivent être construits d'autre corpus, de meilleure qualité, mieux ciblés sur les phénomènes d'évolution et permettant d'évaluer précisément les méthodes présentées dans cet article.

RÉFÉRENCES

- [1] J-F. Allen. Towards a general theory of action and time. *Artificial Intelligence*, 23 :123–154, 1984.
- [2] Nathalie Aussenac-Gilles, Didier Bourigault et Régine Teulier. Analyse comparative de corpus : cas de l'ingénierie des connaissances. In *Actes des 14e journées francophones d'Ingénierie des Connaissances*, pages 67–83. Plate-forme AFIA, 2003.
- [3] Cyril Belica. Analysis of temporal changes in corpora. *International Journal of Corpus Linguistics*, 1(1) :61–73, 1996.
- [4] C.H. Bennett, P. and Ming Li Gacs, M.B. Vitanyi et W.H. Zurek. Information distance. *IEEE Transactions on Information Theory*, 44(4) :1407–1423, 1998.
- [5] Hélène Bestougeff et Gérard Ligozat. *Outils logiques pour le traitement du temps*. Masson, Paris, 1989.
- [6] Éric Brill. A simple rule-based part of speech tagger. In *Proceedings of the Third Conference on Applied Natural Language Processing*, pages 152–155, 1992.
- [7] Maria-Térésa Cabré. *La terminologie, théorie, méthodes et applications*. Armand Colin, Paris, 1998.
- [8] Béatrice Daille. *Approche mixte pour l'extraction de terminologie : statistique lexicale et filtres linguistiques*. Thèse de doctorat en informatique, Université de Paris 7, 1994.
- [9] Béatrice Daille. Variations and application-oriented terminology engineering. *Terminology*, 11(1) :181–197, 2005.
- [10] Ferdinand De Saussure. *Cours de linguistique générale*. Payot, Paris, 1916. Édité en 1969 par Charles Bally et Albert Sechehaye.
- [11] Loïc Depecker. *Entre signe et concept*. Presses Sorbonne Nouvelle, Paris, 2002.
- [12] M-C. Fauvet et P-C. Scholl. Temps et bases de données. Concepts temporels pour la gestion de l'évolution des données. Rapport LGI, IMAG, Grenoble, 1995.

- [13] Benoît Habert, Cécile Fabre et Fabrice Isaac. *De l'écrit au numérique. Constituer, normaliser et exploiter les corpus électroniques*. InterEditions, Paris, 1998.
- [14] Benoît Habert, Adeline Nazarenko et André Salem. *Les linguistiques de corpus*. Armand Colin / Masson, Paris, 1997.
- [15] Christian Jacquemin. Syntagmatic and paradigmatic representations of term variation. In *Proceedings of 37th Annual Meeting of the Association for Computational Linguistics (ACL,99)*, pages 341–348, University of Maryland, 1999.
- [16] André Le Meur et Loïc Depecker. Balnéo : un projet de réseau informatique pour la veille néologique. *Terminologies Nouvelles*, (14) :48–53, 1995.
- [17] Ludovic Lebart et André Salem. *Statistique textuelle*. Dunod, Paris, 1994.
- [18] V. I. Levenshtein. Binary codes capables of correctiong deletions, insertions, and reversals. *Sov. Phys. Dokl.*, 10 :707–710, 1966.
- [19] Bernt Møller. À la recherche d'une terminochronie. *Meta*, XLIII(3), 1998. Revue électronique.
- [20] Denis Monière. Les mots du pouvoir. Cinquante ans de discours inauguraux au Québec (1944-1996). *Lexicometrica*, (spécial "Discours politiques"), 2001. Revue électronique.
- [21] C. Muller. *Principes et méthodes de statistique lexicale*. Champion-Slatkine, Paris-Genève, 1992.
- [22] Fiammetta Namer. Un analyseur flexionnel du français à base de règles. *Traitement automatique des langues*, 41(2) :523–548, 2000.
- [23] Alain Rey. *La terminologie : noms et notions*. Collection Que sais-je ? (1780). Presses Universitaires de France, 2e édition, 1992.
- [24] Sorcha Roche et Lynne Bowker. Système de détection semi-automatique des néologismes. *Terminologies Nouvelles*, (20) :12–16, 1999.
- [25] John Sinclair. Preliminary recommendations on corpus typology. Technical report, EAGLES (Expert Advisory Group on Language Engineering Standards), CEE, 1996.
- [26] Ludovic Tanguy et Nabil Hathout. Webaffix : un outil d'acquisition morphologique dérivationnelle à partir du web. In *Actes de la conférence TALN 2002, Nancy 24-27 juin 2002*, 2002.
- [27] Annie Tartier. *Analyse automatique de l'évolution terminologique : variations et distances*. Thèse de doctorat en informatique, Université de Nantes, 2004.
- [28] R.A. Wagner et M.J. Fischer. The string-to-string correction problem. *ACM*, 21 :168–173, 1974.