Évaluation des outils terminologiques : enjeux, difficultés et propositions

Adeline Nazarenko* — **Haïfa Zargayouna*** — **Olivier Hamon*** ** — **Jonathan van Puymbrouck***

* *Laboratoire d'Informatique de Paris-Nord (CNRS - Univ. Paris 13, UMR 7030)*
99, avenue J.B. Clément, F-93430 Villetaneuse
prenom.nom@lipn.univ-paris13.fr

** *ELDA*
55-57 rue Brillat-Savarin, 75013 Paris
hamon@elda.org

RÉSUMÉ. *Cas particulier parmi les tâches de traitement automatique des langues, l'acquisition terminologique n'a guère fait l'objet d'évaluation systématique jusqu'à présent. Les campagnes qui ont eu lieu sont récentes et limitées. Il est cependant nécessaire de conduire des évaluations pour faire le bilan des recherches passées, mesurer les progrès accomplis et les angles morts. Cet article défend l'idée qu'on peut définir des protocoles d'évaluation comparative même pour des tâches complexes comme la terminologie computationnelle. La méthode proposée s'appuie sur une décomposition des outils d'analyse terminologique en fonctionnalités élémentaires ainsi que sur la définition de mesures de précision et de rappel adaptées aux problèmes terminologiques, à savoir la complexité des produits terminologiques, la dépendance aux applications, le rôle de l'interaction avec l'utilisateur et la variabilité des terminologies de référence.*

ABSTRACT. *In contrast with other NLP subtasks, there has been only few and limited evaluation campaigns for terminology acquisition. It is nevertheless important to assess the progress made, the quality and limitations of terminological tools. This paper argues that it is possible to define evaluation protocols for tasks as complex as computational terminology. Our approach relies on the decomposition into elementary terminological subtasks and on metric definition. We take into account the specificity of computational terminology, the complexity of its outputs, the role of application and user and the absence of well-established gold standard.*

MOTS-CLÉS : *extraction de termes, variation, évaluation, distance terminologique.*

KEYWORDS: *term extraction, term variation, evaluation, terminological distance.*

TAL. Volume 50 – n° 1/2009, pages 257 à 281

1. Introduction

À la croisée de la terminologie traditionnelle et du traitement automatique des langues (TAL), la terminologie computationnelle cherche à construire de manière automatique ou semi-automatique des ressources terminologiques à partir de corpus. Ce domaine de recherche est né au début des années 1990 avec l'essor de l'analyse de corpus mais, du fait de l'accroissement des besoins de gestion de l'information et de leur internationalisation, les ressources terminologiques qui permettent d'indexer les documents, de guider la rédaction de documents techniques, de les traduire, etc. sont de plus en plus nécessaires.

Malgré les années de recul et d'expériences accumulées, il est difficile de se faire une idée claire de l'état de maturité des recherches en terminologie computationnelle. À la différence de nombreux autres pans du TAL, il y a eu peu d'effort collectif pour définir un cadre d'évaluation adapté à ce type de travaux. Les raisons sont diverses. Issue au départ de pays francophones, la terminologie computationnelle n'a pas bénéficié de l'impulsion américaine pour tout ce qui touche aux tâches d'évaluation. De manière générale, c'est une tâche considérée comme moins classique que l'analyse syntaxique, la désambiguïsation ou l'extraction d'information. Il y a aussi d'autres raisons, plus directement liées à la complexité de l'analyse terminologique et à son contexte d'utilisation, sur lesquelles nous reviendrons dans la suite.

Cet article développe la question de l'évaluation des outils d'analyse terminologique, en mettant l'accent sur la terminologie monolingue par opposition à la terminologie multilingue qui constitue un champ de recherche à part entière. Nous considérons l'évaluation des outils eux-mêmes plutôt que des produits terminologiques qu'ils permettent de construire, même si cette distinction est parfois difficile à faire, car on évalue les outils à partir de la qualité de leurs résultats. Nous mettons également l'accent sur les aspects technologiques plutôt que sur les critères ergonomiques ou de performance logicielle, bien que ces deux aspects aient leur importance. Nous pensons qu'il est possible de définir, pour les outils terminologiques, des protocoles d'évaluation comparative qui restent indépendants de toute application particulière.

La première partie de cet article (section 2) souligne les enjeux de l'évaluation des outils d'acquisition terminologique. La section 3 présente les premières expériences qui ont eu lieu dans ce domaine. En s'appuyant sur cet état de l'art et sur l'analyse des difficultés, la section 4 présente nos objectifs en matière d'évaluation des outils terminologiques. Les deux sections suivantes (5 et 6) introduisent nos propositions concernant le découpage de l'analyse terminologique en fonctionnalités élémentaires évaluables indépendamment les unes des autres et la redéfinition des métriques de précision et de rappel. La dernière section (7) apporte une première validation de notre approche : nous exploitons des expériences antérieures d'acquisition de terminologie pour métaévaluer les métriques proposées. Celles-ci sont testées sur l'anglais.

2. Enjeux de l'évaluation en terminologie

Assez marginale au début des années 1990, la terminologie computationnelle s'est développée puis diversifiée avec l'essor des traitements de corpus. Nous nous contentons ici de rappeler les grandes lignes de cette évolution sans mentionner tous des outils terminologiques qui ont été proposés[1].

2.1. *Premiers outils terminologiques*

Les premiers travaux visaient à établir des listes de termes et, pour ce faire, des extracteurs ont été construits. Plusieurs approches ont été adoptées. Certains extracteurs s'appuient sur une analyse morphologique et syntaxique pour identifier les unités pouvant être considérées comme des termes. D'autres outils adoptent au contraire une approche statistique et exploitent des cooccurrences de mots. D'autres extracteurs enfin, combinent les approches syntaxiques et statistiques.

De cette opposition entre approches linguistique et statistique découle le fait que certains extracteurs proposent des listes de termes ordonnées alors que d'autres fournissent des listes « en vrac ». Dans certains cas, on privilégie les termes « correctement formés », dans d'autres la robustesse de l'analyse et la couverture du résultat. Les résultats fournis par les différents systèmes ne sont donc pas entièrement comparables, ce qui a des conséquences en termes d'évaluation, nous y revenons dans la section 6.

2.2. *Diversification des objectifs*

La terminologie computationnelle gagnant en maturité, de nouvelles questions ont vu le jour. Il ne s'agit plus seulement de construire des listes de termes, il faut organiser ces listes en nomenclatures, thesaurus, réseaux sémantiques. Les termes correspondant à des concepts équivalents sont regroupés, des formes canoniques sont distinguées, les termes sont structurés en hiérarchies, les relations sémantiques qu'ils entretiennent sont explicitées, des correspondances sont recherchées entre des termes de différentes langues, etc. Les outils d'acquisition terminologique se complexifient et se dotent ainsi de nouvelles fonctionnalités : calcul de variation terminologique, structuration hiérarchique de terminologies et plus largement calcul de relations sémantiques entre termes et de correspondances multilingues.

Cette diversification rend les outils terminologiques plus difficiles à cerner, à comparer et à évaluer : ils n'ont pas tous les mêmes fonctionnalités et, pire, la définition de ces fonctionnalités varie d'un cas à l'autre.

1. Le lecteur pourra se reporter à des articles de synthèse sur ces sujets (Cabré Castellví *et al.*, 2001 ; Jacquemin *et al.*, 2003 ; Daille *et al.*, 2004).

2.3. *Hétérogénéité des applications*

Il est d'autant plus difficile de faire le bilan des recherches passées en terminologie computationnelle que les ressources terminologiques répondent à des objectifs divers.

Dans les domaines spécialisés, la traduction s'appuie depuis longtemps sur des ressources terminologiques bilingues que l'on cherche aujourd'hui à intégrer dans les systèmes de traduction automatique (Langlais *et al.*, 2004). De nombreux outils visent à établir des correspondances entre les termes d'une langue source et leurs équivalents dans une langue cible à partir de l'analyse de corpus alignés (Kageura *et al.*, 2000).

Le champ traditionnel de l'indexation et de l'analyse documentaire a également été réinvesti par la terminologie computationnelle. Les bases documentaires spécialisées sont indexées sur la base de termes du domaine[2]. Le vocabulaire d'indexation, et donc les termes, sont souvent présentés à l'utilisateur pour l'aider à formuler des requêtes précises (Anick, 2001) et à naviguer dans de gros documents (Wacholder *et al.*, 2001). Cela suppose généralement de sélectionner des formes canoniques « présentables » de termes et de prendre en compte les phénomènes de variation dans le calcul des poids des termes (Jacquemin *et al.*, 1999).

Au-delà de l'indexation, les termes sont utilisés pour faciliter l'interprétation et la visualisation des résultats des outils documentaires (Pratt *et al.*, 1999).

L'analyse terminologique sert encore de support aux tâches visant à conceptualiser un domaine d'activité et à organiser l'ensemble des connaissances qui s'y rattachent. De nombreux travaux s'appuient sur l'analyse terminologique pour amorcer la construction d'ontologies (Meyer *et al.*, 1992 ; Aussenac-Gilles *et al.*, 2004 ; Cimiano, 2006), celles-ci étant à leur tour utilisées pour indexer des documents ou comme référentiel pour la gestion de connaissances (Rinaldi *et al.*, 2005).

La terminologie computationnelle est enfin utilisée comme partie intégrante du processus d'analyse de corpus, dès lors qu'il porte sur les langues de spécialité.

D'une application à l'autre, ce ne sont pas les mêmes propriétés des ressources terminologiques qui sont exploitées : si le regroupement des termes en variantes est utilisé dans beaucoup d'applications, la notion de terme canonique et l'explicitation des relations sémantiques entre termes ne sont pas toujours utiles. Il est important que les termes soient bien formés quand ils doivent être présentés à l'utilisateur mais de simples cooccurrences de mots sont déjà utiles à la traduction automatique.

La diversification des outils d'acquisition terminologique et de leurs applications font qu'il est difficile de cerner le champ de la terminologie computationnelle et d'en mesurer les progrès. Sur le plan de la recherche, le domaine est moins actif qu'il ne l'a été au cours des années 1990 comme si le problème de l'acquisition terminologique était résolu ou, du moins, comme si un palier de performance avait été atteint. Il est d'autant plus important d'établir une cartographie des méthodes et techniques propo-

2. UMLS pour MedLine, par exemple.

sées pour déterminer s'il reste des marges de progression et mettre en correspondance méthodes et types de besoins. Un effort collectif d'évaluation devrait permettre à terme de mieux valoriser les résultats de la terminologie computationnelle.

3. Premières expériences d'évaluation

Même si l'on manque de recul dans ce domaine, il est éclairant d'analyser les expériences d'évaluation qui ont vu le jour en terminologie computationnelle.

3.1. *Les campagnes d'évaluation*

Les expériences les plus notables ont été réalisées dans le cadre de campagnes d'évaluation. Elles visent classiquement à évaluer un ensemble de systèmes, sur une tâche clairement spécifiée et sur un jeu de données commun, en classant les résultats obtenus par les différents systèmes et/ou en les comparant à une référence. Au final, on obtient, sous la forme de mesures, une estimation des performances des systèmes et de leur classement pour la tâche considérée. Ces premières campagnes d'évaluation en terminologie computationnelle présentent des protocoles d'évaluation intéressants.

La campagne d'évaluation NTCIR[3] a débuté en 1999 par un projet d'évaluation de la recherche d'information et de la reconnaissance de termes pour le japonais (*National Center for Science Information Systems, 1999)*. La tâche de reconnaissance de termes (*TEMREC*) était subdivisée en trois sous-tâches : l'extraction de termes, l'extraction de mots-clés et l'analyse des rôles (sujet, action, etc.) des mots-clés. L'évaluation des systèmes s'est faite sur la base d'un ensemble de termes « standard ». Malheureusement, cette tâche n'a pas connu un vif succès et elle a disparu des campagnes NTCIR ultérieures au profit d'autres tâches. Ses organisateurs expliquent notamment cet échec relatif et le nombre limité de participants par l'absence de campagne d'évaluation antérieure (Kageura *et al.*, 2000).

Sans être une campagne à proprement parler, CoRReCT propose un jeu de test et un protocole originaux (Enguehard, 2003). L'objectif est d'évaluer la tâche de reconnaissance de termes en corpus, qui se rapproche de l'indexation contrôlée de documents. Les systèmes prennent en entrée un corpus et une terminologie relevant du même domaine. Ils doivent indexer le corpus avec les termes de la liste fournie, quelle que soit la forme sous laquelle ils apparaissent. CoRRecT a la particularité de reposer sur une construction incrémentale d'un corpus annoté de référence. Au départ le corpus de référence est le corpus brut qui est fourni aux participants, mais il s'enrichit chaque fois qu'un nouveau système participe à l'évaluation. Lorsqu'un nouveau système participant soumet sa liste d'annotations, celles-ci sont confrontées avec l'état courant du corpus de référence. Les nouvelles propositions d'annotations sont évaluées manuellement et celles qui sont validées viennent enrichir le corpus de référence

3. Voir le site http ://research.nii.ac.jp/ntcir/

(phase d'adjudication). Le corpus de référence comporte donc l'union des propositions d'annotations des différents systèmes une fois celles-ci validées. Les résultats du i+1^{e} système sont comparés (en termes de précision et de rappel) avec l'union des résultats validés des i premiers systèmes.

La campagne la plus aboutie est CESART[4] (Mustafa el Hadi *et al.*, 2006). Elle comportait au départ trois tâches (extraction terminologique, indexation contrôlée et extraction de relations) définies de manière peu contraignante[5]. Seule la première a réellement donné lieu à une évaluation, les systèmes candidats à l'évaluation faisant défaut pour les autres[6]. Pour l'extraction, cinq systèmes ont concouru. Cette faiblesse numérique s'explique en partie par le fait que la campagne portait sur le français uniquement mais on peut avancer d'autres raisons liées à l'effort que requiert une participation, à un appel à participation trop peu incisif, au manque de perspective d'amélioration pour les systèmes, etc.

Le protocole élaboré pour la première tâche de CESART est intéressant. Les résultats des extracteurs de termes sont évalués par comparaison avec une liste de termes préétablie par des experts (en pratique, des terminologies préexistantes ont été utilisées). Un corpus d'acquisition relevant du domaine de la terminologie de référence est fourni en entrée des systèmes et ceux-ci doivent en extraire une liste de termes. Cette liste de termes candidats est comparée à la liste de référence. L'une des originalités de CESART a été de considérer la pertinence d'un candidat terme sur une échelle à cinq valeurs plutôt que comme une valeur booléenne. Un terme est considéré comme pertinent s'il apparaît tel que dans la terminologie de référence mais aussi, à un moindre degré, s'il est composé de mots qui relèvent du vocabulaire de la terminologie de référence. Une phase d'adjudication a en outre permis de compléter la référence là où certains termes pertinents manquaient. Cette campagne a également mis en évidence l'hétérogénéité des résultats fournis par les différents systèmes, du point de vue de la longueur des listes de candidats termes fournies. Les extracteurs étaient évalués sur les 10 000 premiers termes proposés mais certains en ont fourni beaucoup plus.

Ces campagnes d'évaluation en terminologie computationnelle souffrent de la comparaison avec d'autres campagnes d'évaluation sur d'autres technologies. Certaines campagnes du programme Technolangue ou les campagnes du NIST ont reçu de nombreux participants et ont été très animées : fort de l'expérience de la précédente campagne Grace[7], la campagne Easy[8] a reçu la participation d'une quinzaine de systèmes. Cela s'explique par l'existence d'une communauté intéressée par le problème

4. Campagne d'évaluation des systèmes d'acquisition des ressources terminologiques. Ce projet Technolangue d'évaluation a été financé par le ministère de la Recherche et de la Technologie et coordonné par l'Université de Lille 3 et ELDA.

5. La première tâche était une « extraction en vue de la création d'une base terminologique » et les extracteurs devaient fournir « au moins 10 000 termes ».

6. Il y a eu respectivement 0 et 1 système participant aux deuxième et troisième tâches.

7. Grammaires et ressources pour les analyseurs de corpus et leur évaluation, http ://www.limsi.fr/TLP/grace/.

8. http ://www.limsi.fr/Recherche/CORVAL/easy/

d'évaluation et qui a compris rapidement l'intérêt d'homogénéiser les protocoles et les entrées/sorties à utiliser sans pour autant chercher à uniformiser les stratégies. Il est vrai aussi que de nombreuses campagnes ayant « réussi » bénéficient de l'apport de campagnes réalisées aux États-Unis (analyse morphosyntaxique, recherche d'information, traduction automatique), ce qui n'est pas le cas de la terminologie.

3.2. *Évaluation au travers des applications*

On a aussi cherché à évaluer l'apport des outils d'acquisition dans les applications.

Une fois extraits, les termes peuvent être exploités pour améliorer la qualité de l'analyse syntaxique, notamment pour réduire les ambiguïtés de rattachements prépositionnels fréquents dans les corpus spécialisés. Cette idée de (Bourigault, 1993) a été reprise et testée par (Aubin *et al.*, 2005) pour adapter un analyseur généraliste à la langue de la biologie. Les premiers résultats montrent que prendre en compte les termes raccourcit le temps d'analyse, simplifie les phrases, diminue beaucoup le nombre d'analyses produites et réduit de 40 % environ le nombre d'analyses erronées.

La terminologie est traditionnellement utilisée pour l'indexation de documents et la recherche d'information. Divers travaux ont cherché à comparer les méthodes d'indexation dans cette perspective : indexation manuelle, indexation libre exploitant des outils d'extraction de termes, indexation contrôlée exploitant une terminologie de référence et un calcul de variation pour retrouver les variantes des termes contrôlées en corpus. Selon les cas, le protocole d'évaluation repose sur les jugements d'indexeurs humains (Daille *et al.*, 2000), sur la similarité des index produits automatiquement et de ceux d'une référence établie ou sur la comparaison des documents retournés sur la base des différents index (Névéol *et al.*, 2006). Dans tous les cas, ces protocoles font appel à une expertise (indexeurs expérimentés ou indexation de MEDLINE).

Les outils d'analyse terminologique étant exploités pour construire des index de fin de livre, leur apport et leur qualité ont été évalués dans ce contexte. L'outil IndDoc (Ait El Mekki *et al.*, 2006) exploite ainsi des outils d'analyse terminologique pour construire des ébauches d'index à partir desquels les indexeurs peuvent travailler pour produire des index finaux. L'approche étant semi-automatique, l'évaluation consiste à mesurer l'apport de l'analyse automatique ou, inversement, la charge de travail qui reste à l'indexeur pour produire un index satisfaisant à partir de l'ébauche d'index produite automatiquement. C'est donc l'écart entre la sortie du système et la sortie retravaillée qui est apprécié : un écart réduit est l'indice d'un faible effort de correction.

Dans les domaines spécialisés, la traduction s'appuie depuis longtemps sur des ressources terminologiques bilingues et on cherche à les intégrer dans les systèmes de traduction automatique. Dans le projet CESTA sur l'évaluation des systèmes de traduction automatique, une tâche originale a permis aux participants d'adapter leur système au domaine spécifique de la santé (Hamon *et al.*, 2008). La comparaison des résultats obtenus avant et après cet enrichissement terminologique met en lumière

l'apport de la terminologie sur les systèmes de traduction automatique[9] : sur cinq systèmes évalués, trois ont obtenu des résultats légèrement meilleurs et deux ont nettement amélioré leurs performances. Une expérience d'évaluation intéressante cherche à mesurer l'apport d'une terminologie bilingue dans l'adaptation d'un système de traduction automatique générique (Langlais *et al.*, 2004).

Même si ces premières expériences sont encourageantes, elles ne donnent pas une vue globale de l'évaluation dans les cadres applicatifs considérés. Pour chaque type d'application, il faudrait intégrer différents outils d'acquisition terminologique dans différents outils applicatifs pour apprécier réellement l'impact des premiers sur les seconds et comparer les méthodes d'acquisition terminologiques en s'affranchissant des modalités de leur intégration.

4. Objectifs

Du fait de leur succès mitigé, ces premières expériences d'évaluation de la terminologie computationnelle n'ont pas encore permis de faire émerger un cadre fédérateur d'évaluation comme il en existe dans d'autres sous-domaines du TAL. Même si ce déficit d'évaluation ne s'explique sans doute pas uniquement par des raisons techniques, c'est sur cet aspect que nous mettons l'accent ici. Nous nous appuyons sur l'analyse de ces difficultés pour proposer un nouveau protocole d'évaluation.

4.1. *Une tâche difficile à définir*

4.1.1. *Complexité de l'objet terminologique*

Même si les outils d'analyse terminologique peuvent paraître simples à évaluer (Popescu-Belis, 2007), il ne faut pas sous-estimer la complexité des produits terminologiques. Les termes sont souvent des unités complexes, composées de plusieurs mots, de longueurs variables et obéissant à des règles de variation multiples en corpus. Ensuite ces termes peuvent être reliés par différents types de relations, de la variation morphologique (*véhicule d'occasion, véhicules d'occasion*) aux relations de synonymie (*voiture d'occasion, automobile d'occasion*) ou d'hyperonymie (*voiture, véhicule d'occasion*). Cette complexité est un frein à l'évaluation. La qualité de la terminologie ne peut se mesurer par un facteur unique, comme on le fait par exemple en reconnaissance vocale avec le taux d'erreur de mots : il faut mesurer à la fois la qualité des termes extraits et celle des relations qu'ils entretiennent. Pour appréhender cette complexité, nous proposons de décomposer l'analyse terminologique en sous-tâches élémentaires et de les évaluer séparément (section 5).

9. Ceci en dépit d'un protocole d'évaluation difficile : la campagne a été réalisée sur une durée relativement courte et sur un corpus d'adaptation de faible volume, environ 20 000 mots.

4.1.2. *Relativité de la référence*

Pour un même domaine, on peut produire plusieurs terminologies différentes qui ne donnent pas la même granularité de description, qui ne reflètent pas le même point de vue et qui peuvent traduire des partis pris terminologiques différents. Ces facteurs d'hétérogénéité sont difficiles à isoler et compliquent l'évaluation : il n'y a généralement pas une référence unique mais, au contraire, un grand choix de terminologies, même pour des terminologues expérimentés (Daille *et al.*, 2000). Pour surmonter le handicap que représente la relativité de toute référence, nous proposons d'ajuster la sortie du système à la référence (voir section 6.1.3). Il faudrait aussi multiplier les évaluations pour tester les systèmes par rapport à différentes références.

4.1.3. *Rôle de l'application*

L'application pour laquelle la terminologie est construite joue un rôle dans les contours de ce que l'on cherche à produire ainsi que sur les critères d'évaluation à prendre en compte. Dans les terminologies bilingues utilisées en traduction automatique, il est essentiel que les termes soient bien formés. C'est également important si les termes sont présentés à un utilisateur. En revanche, des listes de termes bruitées et de simples cooccurrences de mots peuvent suffire s'il s'agit de mesurer le poids des termes dans un document. On remarque aussi que la taille des terminologies utilisées varie d'une application à l'autre.

Même si le poids de l'application est important, nous ne proposons pas de protocole d'évaluation centré sur les applications ici. Nous considérons les évaluations technologiques comme une première étape à la fois plus générique et plus simple. Pour mesurer l'apport de l'acquisition terminologique en termes d'application, il faudra en effet définir un protocole propre à chaque application et comparer les résultats de l'application avec et sans la ressource terminologique tout en veillant à ce que l'intégration de ladite ressource dans le système applicatif n'introduise pas de biais.

4.1.4. *Place de l'interaction*

Le processus d'acquisition terminologique est rarement vu comme un processus entièrement automatique. Le plus souvent les outils terminologiques sont des outils d'aide à l'acquisition qui intègrent la participation du terminologue dans le processus de construction de la ressource. Cette part d'interaction complique l'évaluation de la tâche parce qu'il est difficile de départager ce qui relève du système d'acquisition et la part de travail manuel. Nous défendons cependant l'idée qu'on peut utilement comparer la sortie du système à la sortie validée par l'utilisateur : cela donne une idée du coût de validation qui a été nécessaire pour rendre « acceptable » la ressource produite par le système.

4.2. *Choix d'un protocole comparatif*

La figure 1 présente schématiquement trois scénarios d'évaluation envisageables. Le premier consiste à comparer les sorties du système d'acquisition terminologique à une référence indépendante. Le deuxième évalue l'interaction car il est important de mesurer l'effort fourni par l'expert pour aboutir au produit final à partir de la sortie du système. Cet effort peut être mesuré comme une distance d'édition où l'on compte les opérations élémentaires (suppressions, ajouts et modifications d'éléments terminologiques) permettant de valider la ressource. Le troisième scénario évalue la ressource terminologique indirectement à travers une application (traduction, indexation, etc.) en utilisant les critères d'évaluation propres à l'application considérée.

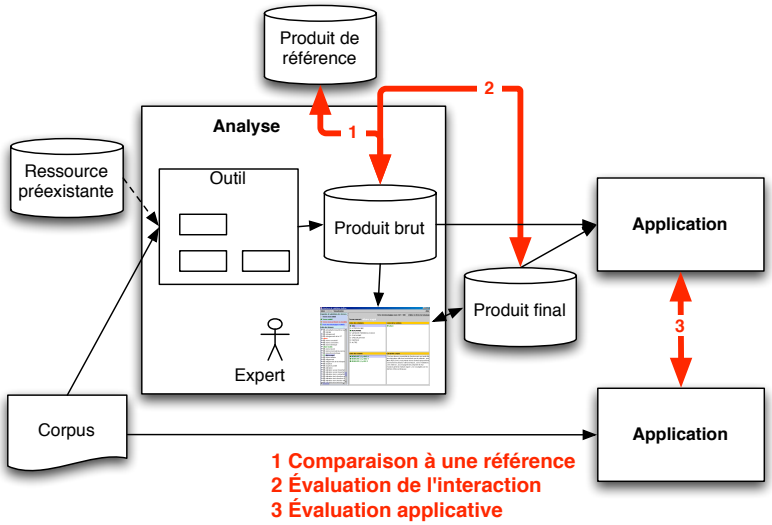

Figure 1. *Différents scénarios d'évaluation : comparaison à une référence (1), évaluation de l'interaction (2) et évaluation applicative (3)*

Nous proposons dans la suite des protocoles et métriques adaptés aux deux premiers scénarios qui relèvent du même schéma, celui d'une évaluation comparative entre ressources : on compare la sortie d'un système avec une terminologie de référence ou la sortie brute d'un système avec cette même sortie validée[10].

10. Ce second cas de figure fait l'économie d'une terminologie de référence mais elle ne permet pas d'évaluer le rappel dans les résultats.

4.3. *Méthodes de construction de la référence*

Toute évaluation comparative présuppose une référence même si celle-ci peut rarement prétendre au statut de *gold standard*. Ce référentiel peut être produit de différentes manières : en réutilisant une ressource existante, comme dans CESART, mais cette dernière risque de n'être que partiellement liée au corpus d'acquisition ; en demandant à un ou plusieurs experts d'acquérir manuellement des termes à partir du même corpus d'acquisition qui est fourni aux systèmes, une solution fiable mais coûteuse ; en fusionnant, comme dans CoRRecT, la validation des résultats de différents systèmes. Cette dernière approche n'est pas exhaustive, mais elle permet de comparer les résultats des différents systèmes entre eux à moindre coût. Nos propositions de métriques ne préjugent pas du choix ou du mode de construction de la référence.

4.4. *Vers de nouvelles métriques*

Les mesures les plus répandues sont le rappel et la précision ainsi que la f-mesure qui les combine. Elles permettent de rendre compte du bruit et du silence des résultats produits par rapport à une référence. Ce sont des mesures faciles à calculer et à interpréter. Elles ont été appliquées à des problèmes très divers.

Néanmoins, ces mesures reposent généralement sur l'hypothèse que la pertinence est une notion binaire (oui/non) alors qu'un terme candidat peut être seulement proche d'un terme de la référence. Dans une expérience d'acquisition terminologique faite à partir de textes anglais de biologie, (Aubin, 2003) rapporte ainsi l'exemple de termes incomplets, comme *core rna*, qui ne sont pas validés en tant que tels mais qui ne sont pas complètement rejetés non plus : ils sont corrigés par l'expert pour former des bons termes, comme *core rna polymerase*, par exemple. C'est pour échapper à cette logique du « tout ou rien » que CESART a défini cinq niveaux de précision.

Comme il n'est pas toujours évident de paramétrer les extracteurs pour spécifier le type de terminologie à produire, il faut adapter leurs sorties avant de mesurer leur rappel et leur précision. Pourquoi en effet pénaliser un extracteur qui produit beaucoup de résultats sous prétexte que la référence choisie est moins prolixe ?

Nous proposons dans ce qui suit des solutions à ces difficultés en décomposant le processus d'acquisition en fonctionnalités élémentaires pour mieux délimiter les objets d'étude (section 5) et en définissant des mesures qui tiennent compte de la spécificité des problèmes terminologiques (section 6).

5. Découpage en tâches élémentaires

Décomposer le processus d'analyse terminologique global en fonctionnalités élémentaires permet d'identifier ce que les outils ont en commun, au-delà de la diversité de leurs méthodes et des applications pour lesquelles ils sont conçus. Ces fonctionnalités ne fournissent pas nécessairement des résultats directement utilisables mais elles

offrent des points de comparaison entre outils. En nous inspirant des campagnes précédentes, nous proposons de distinguer trois fonctionnalités élémentaires génériques et d'évaluer les outils d'acquisition terminologique selon ces trois dimensions.

5.1. *Extraction de termes*

L'extraction terminologique représente la capacité d'un système à établir une liste de termes simples et complexes à partir d'un corpus d'acquisition. La liste plate des termes extraits du corpus est rarement utilisée en tant que telle dans les applications mais la plupart des outils d'acquisition terminologique passent par une étape d'extraction qui constitue donc un bon point d'observation pour l'évaluation. Nous avons pris le parti de dissocier l'extraction terminologique à proprement parler du tri des termes extraits, les critères de pertinence généralement retenus étant surtout dépendants de l'application cible dans laquelle la terminologie doit être exploitée[11].

Spécifier la tâche d'extraction soulève la question de la définition de ce qu'est un terme. L'analyse de corpus montre qu'un même concept peut être désigné de diverses manières, par un nom, par un groupe nominal (terme complexe) et parfois par des verbes ou tournures verbales (par ex. *cliquer*, *fermer une fenêtre*, *installer* en informatique). De ce point de vue, l'évaluation doit être neutre et prendre en compte la diversité des termes et des produits terminologiques, ce qui a un impact sur les métriques utilisées (voir section 6). Le protocole d'évaluation de l'extraction terminologique est simple : un corpus d'acquisition est fourni en entrée au système. Celui-ci produit une liste de termes qu'il faut comparer avec la référence (voir figure 2).

5.2. *Calcul de variation*

Le calcul de variation terminologique consiste à regrouper les familles de termes qui sont des variantes les uns des autres, c'est-à-dire à établir des classes d'équivalence de termes. Pour s'en tenir à des fonctionnalités élémentaires, nous dissocions le fait de construire la classe et le choix d'un terme canonique représentant de la classe qui n'entre pas dans toutes les applications. Évaluer automatiquement le calcul de variation consiste alors à comparer les relations d'équivalence proposées par un système à celles d'un autre système ou d'une référence établie au préalable.

La fonction du calcul de variation est plus difficile à définir que la précédente dans la mesure où la notion de variation pose problème. Certains auteurs distinguent différents niveaux de variation (morphologique, syntaxique ou sémantique) mais ce critère est lié aux méthodes de repérage en corpus plutôt qu'à la nature du résultat. En tenir compte biaiserait l'évaluation. Il faut au contraire considérer la nature de la relation obtenue qui doit refléter une équivalence sémantique. Malheureusement, comme

11. Ils sont souvent liés à la fréquence des candidats termes, à leur répartition dans le corpus d'acquisition, à la typographie, à la position de leurs occurrences dans le texte, etc.

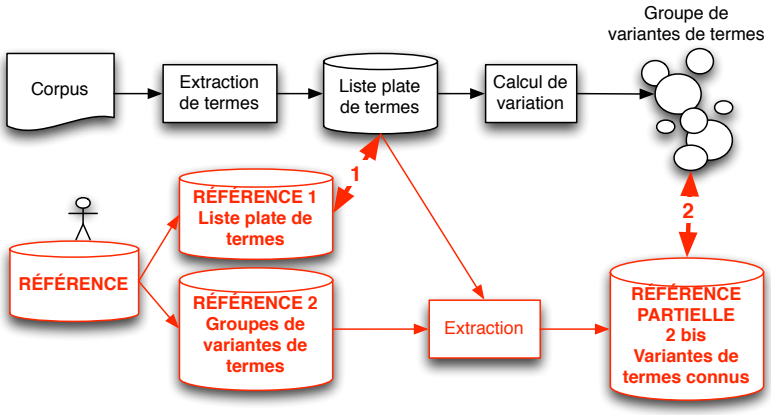

Figure 2. *Protocoles d'évaluation pour l'extraction (1) et le calcul de variation (2). Les deux références 1 et 2 peuvent être extraites de la référence de départ mais la référence 2 ne peut être utilisée telle que pour évaluer les résultats du calcul de variation. Il faut évaluer le calcul de variation sur la seule base des termes connus par le système (référence 2 bis)*

(Daille, 2005) le souligne, la notion d'équivalence sémantique ou conceptuelle (les variantes d'un même terme renvoient au même concept comme *infarctus du myocarde/crise cardiaque*) varie d'une application à l'autre. On considérera donc comme bien formée la classe de termes suivante : *expression de gène, expression génique, gènes exprimés* et *production de gène*, dès lors qu'un terminologue ou la référence pose que *production de gène*, *expression de gène* et l'ensemble des termes de cette classe peuvent être employés l'un pour l'autre.

Le protocole d'évaluation du calcul de variation est également plus complexe que celui de l'extraction (voir figure 2). L'idéal serait de donner une liste plate de termes en entrée aux outils de calcul de variation et de récupérer en sortie une liste de classes d'équivalence de termes à comparer avec une liste de référence de la même forme. Malheureusement, les systèmes calculant la variation s'appuient souvent sur un corpus et ne peuvent pas toujours prendre en entrée une liste de termes. Pour analyser le calcul de variation indépendamment de l'extraction terminologique, nous comparons la sortie du système à un sous-ensemble de la liste de référence comprenant uniquement les relations de variation entre des termes qui sont reconnus par le système.

5.3. *Extraction de relations terminologiques*

L'extraction de relations terminologiques est moins souvent présente dans les outils d'acquisition. Elle consiste à élaborer un réseau de relations sémantiques entre

termes ou classes de termes. La diversité des relations génériques ou spécialisées considérées, leur dépendance au domaine d'application et les différences de granularité des descriptions sémantiques produites font que cette troisième fonction est sans doute la plus difficile à évaluer.

On se rapproche en fait de l'extraction d'information relationnelle[12] où les systèmes annotent un corpus avec les relations sémantiques et où l'étiquetage obtenu est comparé à une annotation de référence. Comme pour le calcul de variation, préserver l'indépendance de l'évaluation de cette tâche par rapport à l'extraction impose de comparer la liste des relations fournies par le système à celles de la référence, en ne prenant en compte que les relations dans lesquelles entrent des termes identifiés par le système. Il faudrait aussi, comme pour l'extraction de termes, tenir compte d'une pertinence graduée, une relation pouvant n'être que « partiellement » pertinente[13].

Réduire la diversité des produits terminologiques à quelques grandes fonctionnalités génériques permet de définir des points de convergence et donc de comparaison entre des outils qui ont souvent des visées et des fonctionnements différents. Même si les résultats de l'extraction, de la variation et de la structuration sont rarement utilisés séparément, ces trois tâches peuvent être évaluées de manière autonome.

6. Métriques d'évaluation

Des métriques doivent être définies pour évaluer les différentes fonctionnalités identifiées et nous mettons ici l'accent sur les deux premières (l'extraction de termes et le calcul de variation) qui nous paraissent prioritaires et plus stabilisées. Les métriques présentées ici sont implémentées dans l'outil Termometer. Nous nous intéressons ici surtout aux principes sur lesquels reposent ces métriques, le détail des formules étant encore susceptible de modification.

Les mesures doivent tenir compte des caractéristiques des produits à évaluer, rester indépendantes des méthodes qu'elles visent à tester et permettre d'évaluer les différentes fonctionnalités indépendamment les unes des autres. Il est aussi important d'avoir des mesures simples, connues et communément admises (Martin *et al.*, 2004) : nous adaptons les mesures classiques de rappel et de précision plutôt que d'en définir de nouvelles[14].

12. Voir l'exemple du Genic Interaction Extraction Challenge http ://genome.jouy.inra.fr/texte/LLLchallenge/

13. Il est moins coûteux de retyper une relation de synonymie en hyperonymie que d'ajouter purement et simplement la relation, par exemple.

14. Rappelons que :

$$precision = \frac{|S \cap R|}{|S|} \qquad rappel = \frac{|S \cap R|}{|R|}$$

où $|S \cap R|$ est le nombre d'éléments pertinents retournés par le système, $|S|$ est le nombre d'éléments retournés par le système et $|R|$ le nombre d'éléments dans la référence.

Les deux tâches d'extraction et de calcul de variation étant complémentaires, elles sont évaluées de manière indépendante et reposent sur des mesures de pertinence distinctes. Le fait qu'un système trouve de bons termes ne doit pas augmenter la pertinence de son calcul de variation et, inversement, une terminologie peut avoir de bonnes variantes et être bruitée. En pratique, un système qui renvoie une liste de termes assortis de variantes est évalué une première fois sur la tâche d'extraction où l'on considère tous les termes proposés indépendamment les uns des autres et une seconde fois sur la tâche de variation où seules les relations de variation sont évaluées.

6.1. *Métriques pour l'extraction de termes*

La tâche d'extraction produit une liste de termes plate (S), dite « de sortie », à comparer à une liste de termes de référence (R). Les mesures de précision et de rappel terminologiques que nous proposons (TP et TR) reposent sur une distance terminologique pour respecter le caractère gradué de la pertinence des termes et sur un ajustement de la sortie à la référence pour tenir compte de la relativité de la référence et s'adapter à la granularité de la description terminologique choisie.

Le comportement global des métriques proposées est le suivant : un système parfait ($S = R$) doit avoir une valeur de qualité maximale ($TP = TR = 1$) ; un système qui renvoie en plus des termes proches des termes de la référence ne doit pas être pénalisé ou faiblement par rapport au système précédent ; un système qui renvoie une liste de non-termes ($S \cap R = \emptyset$) doit avoir la valeur minimale ($TP = TR = 0$). Considérons les cas où $R = \{$*base de données*$\}$ et où on a les listes suivantes en sortie : $S_1 = \{$*base de données, bases de données*$\}$, $S_2 = \{$*bases de données*$\}$ et $S_3 = \{$*base de données, clic droit*$\}$. On souhaite que S_1 et S_2 soient considérées comme de qualité voisine : S_1 retrouve le terme de référence mais y ajoute un deuxième terme redondant ; S_2 ne retrouve qu'un seul terme mais ce n'est pas exactement celui de la référence. S_3 en revanche est de qualité inférieure : le terme surnuméraire est trop éloigné du premier terme pour être considéré comme redondant et il est compté comme bruit.

6.1.1. *Une distance terminologique*

Plusieurs mesures ont été proposées dans la littérature pour apprécier la distance entre les mots. Celles qui reposent sur la comparaison de chaînes de caractères se ramènent toutes plus ou moins à la longueur de la plus grande sous-chaîne commune. La distance d'édition dite de Levenshtein mesure la distance entre deux chaînes de caractères en calculant le coût de transformation de l'une dans l'autre, en fonction des trois opérations élémentaires (ajout, suppression, substitution) qui sont pondérées selon les besoins (Sant, 2004). Les distances « linguistiques » apprécient la distance en fonction du nombre et de la complexité des opérations de transformation linguistiques, notamment morphologiques, qui permettent de dériver un mot d'un autre.

Les termes sont souvent constitués de plusieurs mots ayant leurs règles de variation propres. Deux niveaux interdépendants doivent donc être pris en compte pour mesurer

la distance entre deux termes : celui des mots ou chaînes de caractères (*base* vs *bases*) et celui des termes (*système de fichiers* vs *système de fichiers opérationnel*).

Nous privilégions la distance d'édition qui permet de calculer de manière homogène les distances sur ces deux niveaux comme proposé par (Tartier, 2004). Notre approche diffère cependant de cette dernière parce que nous évitons d'avoir recours à des connaissances linguistiques. Cela permet de simplifier les calculs – ce qui est important étant donné le nombre de distances à calculer en situation réelle – et d'éviter de créer des biais, les connaissances linguistiques introduites pour l'évaluation pouvant aussi être utilisées par certains systèmes pour l'acquisition (règles de dérivation ou étiquetage morphosyntaxique, par exemple). Notre distance terminologique (d_t) repose sur deux distances principales : une distance sur les chaînes de caractères et une distance sur les termes complexes.

La distance sur les chaînes de caractères (notée d_{ch}) est une distance de Levenshtein normalisée. La distance entre deux chaînes est la somme des coûts des opérations élémentaires (insertion, suppression, substitution de caractères) nécessaires pour transformer un mot dans l'autre, en cherchant la séquence d'opérations qui minimise le coût global[15]. Nous divisons ensuite ce coût par le nombre de lettres du plus long mot, afin d'obtenir une distance normalisée comprise entre 0 et 1 et de rendre comparables les distances de différents couples de mots. Si le coût de toutes les opérations est égal à 1 nous obtenons les distances suivantes :

$$d_{ch}(base, bases) = 1/5 = 0,2$$
$$d_{ch}(base, basiques) = 1/8 = 0,125$$
$$d_{ch}(base, relationnelle) = 11/13 = 0,84$$

Quelques exemples ne suffisent pas à valider cette mesure et on trouve des contre-exemples comme *séduction* qui est moins proche de *séduire* ($d_{ch} = 0,44$) que ne l'est *conduire* ($d_{ch} = 0,37$). Cette mesure de distance doit être appréciée globalement à travers l'évaluation d'une liste complète de termes. Nous y revenons dans la section 7.

Nous appliquons le même principe sur les termes complexes ou composés en faisant les opérations de transformation sur les mots qui les composent et non plus sur les caractères. Pour mesurer la distance entre deux termes complexes (d_{tc}), nous nous appuyons cette fois sur le coût des opérations élémentaires (insertion, suppression, substitution de mots) nécessaires pour transformer un terme dans l'autre, en cherchant là encore la séquence d'opérations qui minimise le coût global. Le coût des opérations d'ajout et de suppression est de 1, tandis que le coût de la substitution est égal à la distance de Levenshtein normalisée (d_{ch}) entre les deux mots mis en correspondance.

15. Ce problème d'affectation est résolu par la méthode hongroise (voir http ://en.wikipedia.org/wiki/Hungarian_algorithm) dans Termometer.

Le résultat est normalisé par la longueur du plus long terme. On obtient, par exemple, les distances suivantes :

d_{tc}(*base de données, base de donnée*) $= 0,07$
d_{tc}(*base de données relationnelle, base de données*) $= 0,48$
d_{tc}(*base de donnée, sites Web*) $= 0,85$

Cette mesure permet de rendre compte des phénomènes de permutation fréquents dans certaines langues et de rapprocher des termes comme *expression of gene* et *gene expression*, même si certains de ces rapprochements paraissent abusifs (*base de données* et *données de base* sont deux termes différents). Elle a l'inconvénient de dépendre d'une méthode de segmentation des termes en mots qui est forcément arbitraire.

La distance terminologique se définit comme la moyenne des distances sur les chaînes et sur les termes :

$$d_t(t_1, t_2) = (d_{ch}(t_1, t_2) + d_{tc}(t_1, t_2))/2$$

Les termes sont donc considérés de deux points de vue complémentaires : comme de simples chaînes de caractères et comme des expressions composées de chaînes de caractères. Différentes formules ont été testées expérimentalement mais la moyenne apparaît comme un bon compromis. d_t est une distance normalisée qui varie entre 0 et 1. Les exemples suivants montrent que le facteur d_{ch} atténue la tolérance aux permutations introduite par le facteur d_{tc} et qu'il corrige partiellement l'impact des choix de segmentation :

d_t(*precise gene localization, precise localization of gene*)
$= (10/25 + 1/4)/2 = 0,34$
d_t(*porte folio, portefolios*) $= (1/11 + 3/11)/2 = 0,4$

Cette distance terminologique a le mérite d'être simple à implémenter, facile à interpréter et robuste. Elle ne requiert aucune ressource spécifique et ne dépend pas de la langue. Elle permet d'introduire une pertinence graduée sans préjuger des relations de variation que les systèmes peuvent proposer et qui sont évaluées à part.

6.1.2. *Une mesure de pertinence graduée*

Nous définissons la précision et le rappel terminologique en tenant compte d'une mesure de pertinence graduée. Cette fonction de pertinence $Pert(S, R)$ doit vérifier

$$|S \cap R| \leq Pert(S, R) \leq min(|S|, |R|)$$

et rendre compte d'une proximité globale entre S et R. Elle repose sur les distances terminologiques $d_t(e_s, e_r)$ existant entre les éléments de la sortie S et ceux de la référence R. Pour chaque terme de S, nous retenons le terme de la référence dont il est le plus proche, ce qui permet de définir la pertinence d'un terme comme suit :

$$pert_R(e_s) = \begin{cases} 1 - min_{e_r \in R}(d_t(e_s, e_r)) & \text{si } min_{e_r \in R}(d_t(e_s, e_r)) < \sigma \\ 0 & \text{sinon} \end{cases}$$

où σ est un seuil de distance au-delà duquel nous considérons que deux termes ne sont pas comparables.

6.1.3. *Un ajustement de la sortie à la référence*

Dans la mesure où la référence n'a qu'une valeur relative, il serait artificiel de comparer directement la sortie du système avec la référence. Cela risquerait de trop favoriser le système qui aurait « par hasard » fait les mêmes choix de granularité de description que la référence et les résultats d'évaluation seraient trop dépendants du type de référence adopté. Nous transformons la sortie pour trouver sa correspondance maximale avec la référence, ce qui revient à ajuster la sortie au type de la référence.

Comme plusieurs termes de la sortie peuvent correspondre au même terme de la référence, on peut les considérer en bloc et nous proposons de calculer les mesures de précision et de rappel non pas directement sur S mais sur une partition de S qui est définie relativement à R. Cette partition $\mathcal{P}$(S) est telle que toute partie p de $\mathcal{P}$(S) est soit un ensemble de termes de S qui se rapprochent du même terme de R avec une distance inférieure au seuil σ, soit composée d'un terme singleton :

$$p = \begin{cases} \{e_1, e_2, ..., e_n\} & \text{si } (\exists e_r \in R)((\forall i \in [1,n])(d(e_i, e_r) \leq \sigma)) \\ \{e\} & \text{si } (\nexists e_r \in R)(d(e, e_r) \leq \sigma)\} \end{cases}$$

$$\text{où } e \in S \text{ et } \forall i \in [1,n])(e_i \in S)$$

Nous définissons ensuite la pertinence d'une partie p de $\mathcal{P}(\mathcal{S})$ par rapport à la référence R par la formule suivante[16] :

$$Pert_R(p) = max_{e \in p}(pert_R(e))$$

Les mesures de rappel et de précision terminologiques se définissent comme suit :

$$TP = \frac{Pert(S,R)}{|\mathcal{P}(\mathcal{S})|} = \frac{\sum_{p \in \mathcal{P}(\mathcal{S})} Pert_R(p)}{|\mathcal{P}(\mathcal{S})|}$$

$$TR = \frac{Pert(S,R)}{|R|} = \frac{\sum_{p \in \mathcal{P}(\mathcal{S})} Pert_R(p)}{|R|}$$

Nous pouvons vérifier que dans le cas d'un système parfait *TP = TR = 1*, que *TP* et *TR* tendent vers 0 pour un système qui ne retournerait que du bruit et que *TR* diminue quand la taille de la référence augmente par rapport à celle de la sortie. Le tableau 1 montre ce qu'on obtient comme mesure de précision et de rappel terminologiques pour les trois sorties mentionnées ci-dessus. Considérer la moyenne des distances des termes d'une partie au terme de la référence dont cette partie dépend, plutôt que la distance minimale pourrait permettre de départager S_4 et S_5, mais il faudrait mesurer l'impact d'un tel changement sur des expériences en grandeur nature, ce qui n'a pas encore été fait.

16. À noter que cette pertinence est nulle si p ne contient qu'un élément singleton figurant à une distance supérieure à σ de tout terme de la référence.

Référence		Sortie partitionnée		Pert	TP	TR
R_a	*base de données*	S_1	{base de données, bases de données}	1	1	1
		S_2	{bases de données}	0,93	0,93	0,93
		S_3	{base de données} {clic droit}	1 0	0,5	1
R_b	*cahier des charges*	S_4	{cahier de charge cahier des charges}	1	1	1
		S_5	{cahier des charges}	1	1	1
R_c	*base de données* *bases* *bases de données* *entrepôts de données*	S_6	{base de données}	1	1	0,25

Tableau 1. *Exemples de mesures de TP et TR. Une mesure de pertinence est calculée pour chaque partie de la sortie. Les mesures de pertinence d'une sortie donnée sont ensuite agrégées dans les mesures de TP et de TR*

6.2. *Métriques pour le calcul de variation*

Les métriques utilisées pour évaluer le calcul de variation sont plus simples parce que le protocole d'évaluation défini pour le calcul de variation (*cf.* figure 2) élimine la question du choix des termes à ce stade. La principale difficulté consiste à déterminer la frontière entre ce qui est variation et ce qui ne l'est pas, mais nous supposons cette question résolue dans le choix de la référence.

Les groupes de variantes proposés sont décomposés en listes de couples de termes (t_1, t_2) qui sont la variante l'un de l'autre et nous comparons la liste de couples fournie par le système avec la liste des couples de la référence. Ce sont les mesures traditionnelles de précision et de rappel qui s'appliquent dans ce cas.

7. Métaévaluation des métriques de l'extraction terminologiques

Avant de lancer des expérimentations pour comparer les extracteurs de termes, nous avons voulu valider les métriques proposées pour l'évaluation de l'extraction des termes et donc métaévaluer nos métriques comme cela a pu être fait en traduction automatique (Koehn *et al.*, 2006).

Nous avons pour cela réutilisé des données existantes construites de manière indépendante. Les deux séries de tests que nous présentons portent sur des données fournies par le laboratoire Mathématique Informatique et Génome (MIG) de l'INRA. Même si ces matériaux, qui ont été constitués pour d'autres expérimentations, com-

	Précision	Rappel	F-mesure	TP	TR	F-mesure
Référence	1,0	1,0	1,0	1,0	1,0	1,0
Acabit	0,71	0,42	0,52	0,95	0,48	0,63
Syntex	0,77	0,68	0,72	0,94	0,70	0,80
Nomino	0,76	0,28	0,40	0,95	0,34	0,50

Tableau 2. *Résultats des évaluations sur les sorties des 3 systèmes*

portent des biais, ils nous ont permi de valider la robustesse des métriques proposées et de vérifier leur adéquation à nos spécifications de départ.

7.1. *Comparaison par rapport aux mesures classiques*

Nous disposons des résultats d'un travail de comparaison de différents extracteurs de termes effectué dans le cadre du projet Caderige[17] (Aubin, 2003). Le protocole de test est le suivant :

– le corpus recouvre des domaines comme la biologie moléculaire et la génomique en anglais. Il est constitué d'environ 405 000 mots ;

– on dispose des sorties de trois extracteurs de termes sur ce corpus. Seuls les termes ayant plus de 20 occurrences ont été retenus pour limiter le travail d'évaluation de l'expert et par souci d'efficacité. On a ainsi 194, 307 et 456 candidats termes respectivement fournis par Nomino (David *et al.*, 1990), Acabit (Daille, 1994) et Syntex (Bourigault *et al.*, 2000) ;

– la référence contient 514 termes, elle a été construite en demandant à un expert de valider les productions des différents extracteurs. L'expert peut juger un terme incomplet et proposer une forme complète. Certaines incohérences (5 % des termes) ont été relevées dans le jugement de l'expert qui a accepté certains termes en sortie d'un extracteur alors qu'il les a refusés quand ils étaient proposés par un autre système. Ces incohérences ont été lissées pour aboutir à une référence cohérente. La validation des termes s'est effectuée hors contexte (sans retour vers le corpus), ce qui peut expliquer en partie cette variation intra-annotateur.

Le tableau 2 donne les résultats de l'évaluation des sorties des systèmes par rapport à la référence. Comme attendu, les valeurs de précision et rappel terminologiques (TP et TR) suivent les mêmes tendances que celles des mesures de précision et rappel classiques, mais elles sont plus élevées, preuve que nos mesures permettent de rendre compte d'une certaine approximation de la référence. Les écarts de f-mesure sont significatifs, de l'ordre de 10 points entre nos métriques et les métriques classiques.

Le partitionnement de la sortie conduit en majorité à regrouper des variantes d'ordre morphologique (singulier/pluriel) ou typographique (majuscule/minuscule).

17. http ://caderige.imag.fr

Comme (Aubin, 2003) indique que 8,5 % des candidats termes ont été jugés incomplets, nous avons analysé ces cas plus précisément. On observe que les termes incomplets fournis par les systèmes sont rapprochés des termes complets que l'expert a ajoutés dans la référence. *acid residue* est ainsi rapproché du terme *amino acid residues* avec une distance de 0,35. Dans ce cas particulier, lors du partitionnement de la sortie, quatre termes sont en réalité rapprochés de *amino acid residues* :

```
référence :  amino acid residues
candidat : acid residue (distance = 0,35)
candidat : amino acid residue (distance = 0,05)
candidat : acid residues (distance = 0,29)
candidat : amino acid residues (distance = 0,0)
```

7.2. *Comportement à grande échelle des métriques*

La deuxième expérimentation porte sur la validation des résultats d'un seul système d'extraction de termes (YaTeA, développé au LIPN (Aubin *et al.*, 2006)) dans le cadre du projet Epipagri[18]. Le protocole de test est le suivant :

– le corpus recouvre le domaine des brevets sur les agrobiotechs en anglais ;

– la sortie de YaTeA est évaluée telle quelle. Aucun filtrage ni tri n'a été effectué. YaTeA a extrait près de 4 200 candidats pour ce corpus ;

– la référence contient 1 988 termes. Elle est formée des résultats de la validation de la sortie de YaTeA par deux experts. Les variations interannotateurs ont été résolues par consensus, elles étaient de l'ordre de 11%. Par construction, cette référence n'est pas pertinente pour le calcul du rappel, les experts s'étant contentés de valider ou d'invalider les termes proposés par YaTeA, sans en ajouter de nouveaux. Nous présentons les résultats en termes de précision terminologique uniquement.

L'intérêt de cette expérience est son effet « d'échelle ». Elle permet d'analyser globalement le comportement de nos métriques sur un échantillon de taille importante. Pour cette expérimentation, Termometer effectue environ 8×10^6 calculs de distance terminologique en moins de 10 minutes sur un PC standard.

Les résultats présentés dans la figure 3 montrent le lien entre les valeurs du seuil σ (le seuil de distance au-dessus duquel un candidat terme n'est pas regroupé avec d'autres) et la précision terminologique (fig. 3 a.). Quand $\sigma = 0$, il n'y a aucun regroupement (la précision terminologique et la précision classique sont identiques) mais la précision terminologique augmente avec le seuil. La valeur de σ a un effet direct sur la taille de la partition de la sortie (fig. 3 b.) : plus la valeur du seuil est élevée plus grand est le nombre de termes alignés à la référence (nombre de termes regroupés). Quand σ a la valeur maximale, cela revient à aligner tous les candidats termes et la précision terminologique tend vers 1. Les allures des courbes laissent penser qu'il

18. http ://www.epipagri.org/

est possible de déterminer ce seuil automatiquement autour des points d'inflexion des courbes (entre 0,4 et 0,5 dans notre cas).

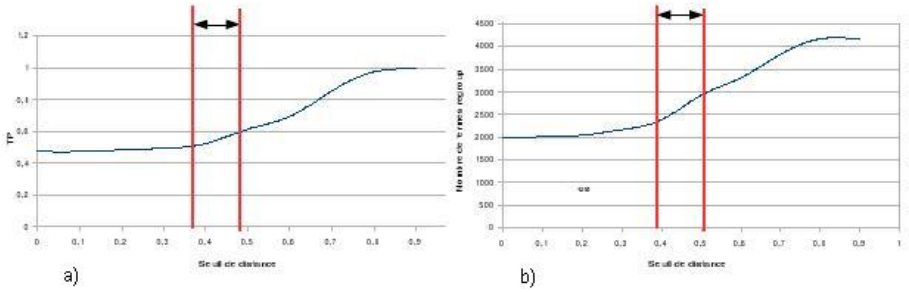

Figure 3. *Courbes de précision terminologique (TP) en fonction des variations du seuil de distance (a) et du nombre de termes alignés à la référence par rapport aux différentes valeurs de ce seuil (b)*

Nous avons également vérifié que nos mesures reflètent la qualité relative des différentes listes de termes fournies par cette expérience. Nous avons considéré la sortie brute (S_b), la sortie validée par le premier expert (V_1), la sortie validée par le deuxième expert (V_2) et la sortie validée par les deux experts après discussion (V_{12}). Nous avons calculé la précision terminologique de chacune de ces listes en prenant V_{12} comme référence. Les résultats figurent dans le tableau 3.

	S_b	V_1	V_2	V_{12}
TP	0,55	0,91	0,97	1,0

Tableau 3. *Précision terminologique sur des sorties de qualité graduée, σ = 0,4*

On voit que le jugement du deuxième expert est plus proche du consensus représenté par V_{12} que celui du premier mais on peut surtout vérifier qu'on obtient les relations d'ordre attendues en termes de précision terminologique entre les différentes listes de termes :

$$TP(S_b) < TP(V_1) < TP(R) \text{ et } TP(S_b) < TP(V_2) < TP(R)$$

Le comportement global des métriques que nous avons proposées paraît cohérent et en ligne avec notre projet initial de renforcer la précision des extracteurs qui fournissent des listes partiellement redondantes.

Ces expériences permettent de valider ici les grands principes de notre méthode d'évaluation de la terminologie, même s'il faut aller plus loin dans l'expérimentation pour valider les métriques proposées et les paramétrer dans le détail, notamment pour la détermination du seuil.

8. Conclusion

Après une quinzaine d'années de recherche sur les méthodes d'acquisition de ressources terminologiques à partir de corpus textuels, il est important de faire le bilan des progrès accomplis et de montrer l'apport de la terminologie computationnelle à l'analyse des corpus spécialisés. Pour ce faire, le travail d'évaluation est essentiel. La faible culture de l'évaluation en terminologie computationnelle surprend en contraste avec d'autres sous-domaines du TAL. Le fait que les outils terminologiques soient souvent des outils d'aide au terminologue, qu'il n'existe pas de référence stable en matière de produits terminologiques et que la nature des terminologies à construire dépende en grande partie des applications qui les exploitent, constituent autant de handicaps pour la définition de protocoles d'évaluation.

Nous défendons néanmoins l'idée que l'on peut définir des protocoles d'évaluation comparative, en prenant appui sur le savoir-faire d'autres domaines du TAL et sur les premières expériences d'évaluation en terminologie. Nous proposons une décomposition de l'analyse terminologique en tâches élémentaires qui peuvent être évaluées indépendamment les unes des autres. Pour la première de ces tâches, l'extraction de termes, nous définissons des métriques d'évaluation terminologiques. Les mesures de précision et le rappel terminologiques sont simples à calculer et à interpréter comme les mesures classiques dont elles sont inspirées. Elles en diffèrent néanmoins sur deux points. Elles reposent sur une mesure de pertinence graduelle plutôt que booléenne pour rendre compte du fait qu'un terme peut être seulement « partiellement pertinent ». Elles sont calculées sur la sortie d'un système une fois que celle-ci a été transformée pour être ajustée à la taille et à la granularité de la référence. L'idée ici est de tenir compte de la relativité de la référence dans l'évaluation des systèmes. Dans le cas du calcul de variation, le protocole est plus complexe mais les métriques plus simples que pour l'extraction.

Les premières expériences de métaévaluation de l'extraction de termes montrent le bon comportement global des mesures que nous avons définies. Les expériences à mener maintenant sont des évaluations en grandeur nature. Le protocole d'évaluation n'est qu'un point de départ. Seule l'accumulation des évaluations faites sur des bases comparables (sur les mêmes tâches et avec les mêmes métriques) peut permettre de faire un état des lieux global de la terminologie computationnelle. Il serait utile pour cela que les données de test soient rendues publiques pour servir de banc d'essai.

Remerciements

Ce travail a été partiellement réalisé dans le cadre du programme Quaero, financé par OSEO, l'agence française pour l'innovation. Les auteurs remercient M. El Moueddeb qui a donné son nom à Termometer ainsi que les collègues avec qui ils ont discuté des questions abordées ici, au sein du groupe TIA (www.tia.org) et du projet CTC du programme Quaero. Ils remercient particulièrement S. Aubin (INRA-MIG) pour le matériau terminologique et les résultats de validation utilisés pour la métaévaluation.

9. Bibliographie

Ait El Mekki T., Nazarenko A., « An application-oriented terminology evaluation : the case of back-of-the-book indexes », *in* R. C. *et al.* (ed.), *Proc. of the LREC Workshop "Terminology Design : Quality Criteria and Evaluation Methods"*, Genova, Italy, p. 18-21, May, 2006.

Anick P. G., « The automatic construction of faceted terminological feedback for interactive document retrieval », *in* D. Bourgault, C. Jacquemin, M. L'Homme (eds), *Recent Advances in Computational Terminology*, John Benjamins, Amsterdam, p. 29-52, 2001.

Aubin S., Comparaison de termes extraits par Acabit, Nomino, Syntex de fréquences supérieures ou égales à 20, Livrable n° 3.2, Projet ExtraPloDocs, INRA-MIG, 2003.

Aubin S., Hamon T., « Improving Term Extraction with Terminological Resources », *in* T. Salakoski, F. Ginter, S. Pyysalo, T. Pahikkala (eds), *Proc. of the Advances in Natural Language Processing (Proc. of FinTAL'06*, LNAI 4139, Springer, p. 380-387, 2006.

Aubin S., Nazarenko A., Nédellec C., « Adapting a General Parser to a Sublanguage », *Proc. of the Int. Conf. on Recent Advances in Natural Language Processing (RANLP'05)*, Borovets, Bulgaria, p. 89-93, 2005.

Aussenac-Gilles N., Biébow B., Szulman S., « Modélisation du domaine par une méthode fondée sur l'analyse de corpus », *in* R. Teulier, J. Charlet, P. Tchounikine (eds), *Ingénirie des Connaisances*, Eyrolles, 2004.

Bourigault D., « An Endogenous Corpus-Based method for Structural Noun Phrase Disambiguation », *Proc. of the 6th European Chapter of ACL*, 1993.

Bourigault D., Fabre C., « Approche linguistique pour l'analyse syntaxique de corpus », *Cahiers de Grammaires*, vol. 25, p. 131-151, 2000.

Cabré Castellví M., Estopà R., Vivaldi Palatresi J., « Automatic Term Detection : A Review of Current Systems », *in* D. Bourigault, C. Jacquemin, M.-C. L'Homme (eds), *Recent Advances in Computational Terminology*, John Benjamins, Amsterdam, 2001.

Cimiano P., *Ontology Learning and Population from Text*, Springer, 2006.

Daille B., Approche mixte pour l'extraction de terminologie : statistique lexicale et filtres linguistiques, Thèse d'informatique, Université Paris VII, 1994.

Daille B., « Variations and application-oriented terminology engineering », *Terminology*, vol. 11, n° 1, p. 181-197, 2005.

Daille B., Kageura K., Nakagawa H., Chien L.-F. (eds), *Terminology. Special issue on Recent Trends in Computational Terminology*, vol. 10, John Benjamins, 2004.

Daille B., Royauté J., Polenco X., « Évaluation d'une plate-forme d'indexation des termes complexes », *Traitement Automatique des Langues*, vol. 41, n° 2, p. 396-422, 2000.

David S., Plante P., « De la nécessité d'une approche morpho-syntaxique en analyse de textes », *revue ICO Québec*, vol. 2, n° 3, p. 140-155, 1990.

Enguehard C., « CoRRecT : Démarche coopérative pour l'évaluation de systèmes de reconnaissance de termes », *Actes de la 10^{e} conférence annuelle sur le Traitement Automatique des Langues (TALN 2003)*, Nancy, p. 339-345, 2003.

Hamon O., Popescu-Belis A., Hartley A., Mustafa El Hadi W., Rajman M., « CESTA : Campagne d'Évaluation des Systèmes de Traduction Automatique », *in* S. Chaudiron, K. Choukri (eds), *L'évaluation des technologies de traitement de la langue*, Hermès, Lavoisier, Paris, p. 93-116, 2008.

Jacquemin C., Bourigault D., « Term Extraction and Automatic Indexing », *in* R. Mitkov (ed.), *Handbook of Computational Linguistics*, Oxford Univ. Press, chapter 19, p. 599-615, 2003.

Jacquemin C., Tzoukermann E., « NLP for term variant extraction : synergy between morphology, lexicon, ans syntax », *in* T. Strzalkowsko (ed.), *Natural Language Information Retrieval*, Kluwer, Boston, MA, p. 25-74, 1999.

Kageura K., Fukushima T., Kando N., Okumura M., Sekine S., Kuriyama K., Takeuchi K., Yoshioka M., Koyama T., Isahara H., « IR/IE/Summarisation Evaluation Projects in Japan », *LREC 2000 Workshop on Using Evaluation within HLT Programs*, p. 19-22, 2000.

Koehn P., Bertoldi N., Bojar O., Callison-Burch C., Constantin A., Cowan B., Dyer C., Federico M., Herbst E., Hoang H., Moran C., Shen W., Zens R., « Factored translation models », *in* J. H. University (ed.), *CLSP Summer Workshop Final Report WS-2006*, 2006.

Langlais P., Carl M., « General-purpose statistical translation engine and domain specific texts : Would it work ? », *Terminology*, vol. 10, n° 1, p. 131-152, 2004.

Martin A. F., Garofolo J. S., Fiscus J. C., Le A. N., Palett D. S., ang Gregory A. Sanders M. A. P., « NIST Language Technology Evaluation Cookbook », *Proc. of 4th Int. Conf. on Language Resources and Evaluation (LREC 2004)*, 2004.

Meyer I., Skuce D., Bowker L., Eck K., « Towards a new generation of terminological resources : an experiment in building a terminological knowledge base », *Proc. of the 15th Int. Conf. on Computational Linguistics (COLING'92)*, Nantes, France, p. 956-960, 1992.

Mustafa el Hadi W., Timimi I., Dabbadie M., Choukri K., Hamon O., Chiao Y., « Terminological Resources Acquisition Tools : Toward a User-oriented Evaluation Model », *Proc. of the Language Resources and Evaluation Conf. (LREC'06)*, Genova, Italy, p. 945-948, 2006.

National Center for Science Information Systems (ed.), *Proc. of the 1st NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*, 1999.

Névéol A., Zeng K., Bodenreider O., « Besides precision & recall : Exploring alternative approaches to evaluating an automatic indexing tool for MEDLINE », *Proc. of the AMIA Annual Symposium*, p. 589-593, 2006.

Popescu-Belis A., « Le rôle des métriques d'évaluation dans le processus de recherche en TAL », *Traitement Automatique des Langues*, vol. 8, n° 1, p. 67-91, 2007.

Pratt W., Hearst M. A., Fagan L. M., « A Knowledge-Based Approach to Organizing Retrieved Documents », *Proc. of the AAAI Conference*, p. 80-85, 1999.

Rinaldi F., Yuste E., Schneider G., Hess M., Roussel D., « Ontology Learning from Text : Methods, Evaluation and Applications », *in* B. M. Paul Buitelaar, Philipp Cimiano (ed.), *Exploiting Technical Terminology for Knowledge Management*, IOS Press, p. 140-154, 2005.

Sant P. M., « Levenshtein Distance », , in *Dictionary of Algorithms and Data Structures* [http ://www.itl.nist.gov/div897/sqg/dads/HTML/rootedtree.html (accessed 2 April 2009)], Paul E. Black, ed., U.S. National Institute of Standards and Technology, April, 2004.

Tartier A., Analyse automatique de l'évolution terminologique : variations et distances, PhD thesis, Université de Nantes, 2004.

Wacholder N., Evans D., Klavans J., « Automatic identification and organization of index terms for interactive browsing », *Proc. of 1st ACM/IEEE-CS Joint Conf. on Digital Libraries*, Roanoke, VA, p. 126-134, june, 2001.