Évaluation des systèmes d'acquisition de terminologie : nouvelles pratiques, nouvelles métriques

Ismaïl Timimi

Lab. GERIICO – Université Charles de Gaulle Lille3

BP 60 149 – 59 653 Villeneuve d'Ascq cedex – France

Abstract

This article is part of the ongoing "Technolangue" research project devoted to the evaluation of tools for the automatic treatment of oral and written corporas. The first section offers a reflection on the evaluation practice as a method of observation strongly linked to socio-technological processes of innovation. We attempt to establish relevant criteria for the appreciation of terminological ressources acquisition tools, and to demonstrate how these criteria remain applicable only within the context of a specific information need or use. Our concluding discussion seeks to resituate current debate on the evaluation within the context of the interdependence between innovation and practice.

Résumé

Cet article s'inscrit dans les travaux du projet Technolangue sur l'évaluation des outils de traitement automatique des corpus écrits et oraux. Nous nous intéressons dans une première partie à une réflexion sur la pratique de *l'évaluation* en tant que méthodologie d'observation fortement liée aux processus d'innovation socio-technologique. On tentera en particulier de déterminer quels sont les critères d'appréciation d'outils d'acquisition de ressources terminologiques, mais aussi, comment ces critères ne peuvent être déterminants que pour un besoin informationnel et d'usage particulier. Notre questionnement final nous permettra de refocaliser les débats sur la *culture de l'évaluation* comme résultat de l'interdépendance entre l'innovation et les pratiques.

Mots-clés : paradigme d'évaluation, extraction de terminologie, ressources terminologiques, indexation, recherche d'information, référentiel d'appariement.

Avant-propos

Dans ce document, sont relevés les différents paramètres de l'évaluation dans le cadre du projet CESART (*Campagne d'Évaluation des Systèmes d'Acquisition des Ressources Terminologiques*). Certains points sont extraits du protocole d'évaluation du projet Aupelf-Arc A3 mais enrichis de nouvelles réflexions tenant compte de la particularité des nouveaux outils participants (Amar *et al.*, 2001 ; Mustafa, 2001).

1. Philosophie de l'évaluation

Peu connue comme activité normalisée, l'évaluation des systèmes des traitements de corpus est depuis quelques années au centre des préoccupations des institutions politiques, scientifiques et industrielles. On distingue plusieurs types d'évaluation suivant les attentes des acteurs (concepteur, développeur, distributeur, utilisateur final).

1.1. Le relativisme de l'évaluation

Si l'évaluation est un exercice d'appréciation des performances d'un système assorti d'un indice de satisfaction, cet indice ne peut être calculé que si l'observation est effectuée par rapport à des besoins et des référentiels de type intrinsèque ou extrinsèque. La diversité des besoins et des référentiels génère plusieurs types et degrés d'évaluation.

Dans une évaluation dite *de progression (verticale)*, le système est comparé à ses versions antérieures pour une tâche déterminée, en vue d'une étude diachronique de ses performances. C'est une démarche très courante dans les activités de conception et de développement de systèmes.

Une deuxième démarche d'évaluation, dite *d'appariement (transversale)*, consiste à comparer les performance d'un système par rapport soit à d'autres systèmes conçus pour des applications similaires, soit à des résultats (besoins) prédéfinis, établis manuellement ou autrement, et surtout validés.

Dans une autre optique d'évaluation, dite *de diagnostic*, l'évaluateur expert cherche à déterminer à partir d'une série de tests les sources de performance ou non d'un système conçu pour une tâche précise. Ce mode est, lui aussi, orienté conception dans la mesure où ces tests de diagnostic permettent de développer par progression les performances d'un système.

Quel que soit le mode adopté, l'évaluation de systèmes de traitement automatique de corpus[1], ne peut être exercée que dans une démarche comparative... par rapport à d'autres systèmes, à des référentiels préétablis, à une application bien déterminée.

1.2. Les facettes de l'évaluation

Dans le cas particulier des outils d'ingénierie linguistique, on peut relever deux démarches dans la méthodologie d'évaluation, qui ne sont pas forcément exclusives. Il s'agit de l'évaluation avec *interface statique* par opposition à *l'interface dynamique* d'une part ; et de l'évaluation de type *boîte noire* par opposition à la *boîte transparente* d'autre part.

L'évaluation d'un système avec *interface statique* consiste à juger ses performances, sans faire appel à des interventions ou à des enrichissements extérieurs. À l'inverse, l'évaluation avec *interface dynamique* permet de calculer l'amélioration des performances d'un système suite à une intégration de ressources extérieures (par enrichissement terminologique par exemple). (Chaudiron, 2001)

Parallèlement, l'activité évaluative dans sa forme classique peut être menée sur le concept de la *boîte noire*. Elle porte sur le jugement des performances globales du système à partir seulement des ressources fournies en entrée (Input) et des résultats produits en sortie (Output), sans examiner le traitement intermédiaire des données effectué par les divers modules du système. À l'opposé, l'évaluation orientée *boîte transparente* s'intéresse à l'étude du fonctionnement interne du système à travers ses différents modules et prétraitements. Elle rejoint dans cet aspect *l'évaluation de diagnostic* précitée.

1.3. L'évaluation dans le protocole de Cesart

Les différentes formes d'évaluation (citées brièvement ci-dessus) sont combinables et adaptables suivant les contextes et les enjeux de la campagne d'évaluation d'une part et la typologie des systèmes participant d'autre part. Dans le cadre du projet Cesart, nous avons

[1]Le cadre de notre objet d'étude ici (Technolangue).

abandonné certaines formes d'évaluation telles que *l'évaluation de progression et de diagnostic*, ainsi que *l'évaluation à la boîte transparente.*

L*'évaluation de progression* a été abandonnée dans la mesure où nous ne pourrions disposer de tous les logiciels, et encore moins de leurs versions antérieures pour pouvoir étudier l'évolution diachronique de leurs performances. Nous avons abandonné également l'évaluation sur le principe de la *boîte transparente*, car il s'agit d'un mécanisme difficile à mettre en place, il requiert une connaissance des processus internes et des fondements théoriques de chacun des systèmes participant au projet. Il réclame l'accès à l'architecture et à la stratégie du système, ce qui risque d'être compromettant lorsque l'évaluateur est un intervenant extérieur (Cavazza, 1993). Pour les mêmes raisons, il ne nous était pas possible d'adopter *l'évaluation de diagnostic*, qui est aussi orientée conception et proche dès lors de l'évaluation de type *boîte transparente*.

Nous avons adopté dans le projet Cesart, le principe de l'évaluation *boîte noire.* Ce choix est justifié du fait qu'il s'agit d'une démarche d'expertise facile à mettre en œuvre, dans un consortium composé d'universitaires et d'industriels et pose le moins de problèmes méthodologiques. Sans nécessiter l'accès au fonctionnement interne des systèmes, elle permet une étude comparative malgré la différence des architectures employées (Cavazza, 1993), (Sparck-Jones *et al.*, 1996).

Afin de combler les limites de l'évaluation *boîte noire*[2], nous avons renforcé le protocole de nouveaux critères d'appréciation et consignes de jugement, dont une partie est inspirée des autres formes d'évaluation.

- Nous avons opté pour une *évaluation d'appariement*, grâce à des métriques quantitatives calculées à partir d'un algorithme d'appariement des systèmes à des référentiels terminologiques préétablis. L'évaluation qualitative est aussi envisagée grâce à des classifications de pertinence proposées par des experts humains, sur des critères fixés préalablement.
- L'évaluation est en adéquation à des contextes prédéterminés tenant compte des besoins de l'utilisateur (*interface dynamique*) et des domaines d'applications (usages).
- Enfin, pour étudier l'extensibilité des systèmes, leur réserve de performance et la possibilité de leurs maintenances (Cavazza, 1993), une partie du principe de l'évaluation avec *interface dynamique* est introduite dans le projet Cesart. Un questionnaire des prétraitements (sous forme de tableau de bord) est pris en considération dans le calcul par les experts des coûts[3] de l'usage prévu. Ce coût dépend des besoins en ressources d'enrichissement internes et externes, des performances requises, du nombre et de nature d'interventions de l'utilisateur du système et du temps de traitement.

2. Projet Cesart : un cadre de réflexion et de validation

La campagne Cesart s'inscrit dans le cadre du Projet Technolangue Chapitre Evalda, co-organisée par le laboratoire Cersates de l'université Lille3 et Elda. Elle consiste à élaborer un protocole *normalisé* pour l'évaluation de systèmes d'acquisition de ressources terminologiques.

[2]Par exemple, une des limites de l'évaluation *boîte noire* est de ne pas prendre en compte les choix et les renseignements apportés au système par son utilisateur dans les étapes préliminaires.
[3]En conformité avec la norme ISO 9126 (King, 1996).

2.1. Présentation des systèmes participants

Les systèmes dont il est question dans ce projet sont issus du milieu universitaire et industriel. Ils proposent des niveaux de traitements et des applications assez variés dans lesquels les termes occupent une place centrale. Leur point commun est donc d'être fondé sur le traitement des termes et des connaissances terminologiques. Les systèmes participants actuellement sont : *IDE/XTS* (TEMIS), *Lexter* (EDF R&D), *WorldTrek* (EDF R&D), *Termic* (CEA), *Termos* (RALI), *SeekJava* (LALICC, Paris 4), *SynoTerm* et Terminae (LIPN, Paris 13), *TermWatch* (LITA, Univ. Metz et ERSICOM, Lyon 3).

Au-delà des différences dans leurs modèles théoriques et leurs architectures, nous avons réparti ces systèmes en trois catégories qui sont davantage liées aux types de tâches pour lesquelles un protocole d'évaluation en concertation pouvait être mis en place[4].

- 4 Extracteurs de Termes : IDE/XTS, Lexter, Termic, Termos
- 4 Extracteurs de Relations syntaxiques : Lexter, Termos, Termic, TermWatch[5]
- 3 Extracteurs de Relations sémantiques : IDE/XTS, SeekJava, SynoTerm
- 2 Éditeurs : Terminae (d'ontologies), WorldTrek (d'interfaces)

Nous traitons à part cette dernière catégorie composée des systèmes qui prennent en entrée les résultats fournis par des systèmes précédents.

Certains systèmes extracteurs de relations produisent également des « classes de termes » reliées. L'évaluation consisterait dans ce cas à l'observation de la cohésion ou non des classes produites et leur adéquation à l'application préétablie.

Tout système *orphelin,* qui représente une singularité dans sa catégorie ou par rapport à une application envisagée fait l'objet d'une évaluation particulière (non comparative à d'autres systèmes). On pense notamment au système *Terminae,* éditeur d'ontologies et *TermWatch* pour son application principale, la production de classes de thématiques.

[4]Ces catégories ne résument donc pas à elles seules les traitements et les applications cibles de tous ces systèmes.
[5] Ce dernier n'est pas forcément un extracteur de relations mais plutôt un outil dédié à la fouille de textes spécialisés dans le but d'extraire des informations pouvant être exploitées dans un contexte de veille.

		Lexter	Termos	Termic	TermWatch	IDE/XTS	SynoTerm	SeekJava	Lexter	Terminae
Extracteurs de	Termes	X	X	X		X				
	Concepts									X
Extracteurs de	Relations syntaxiques	X	X	X	X					
	Relations sémantiques					X	X	X		
Extracteurs de	Classes thématiques				X					
	Classes morpho-syntaxiques	X	X	X	X					
	Classes sémantiques						X			
Éditeurs	Éditeurs d'ontologies									X
	Structuration de termes					X	X		X	

Répartition des outils participant à la campagne

2.2. Évaluation orientée applications

Si dans certains projets d'évaluation, les outils participants ont des applications clairement identifiées (résumé, traduction, question-réponse...), et sur lesquelles porte l'activité d'évaluation proprement dite, les outils participant au projet Cesart présentent, de par leur fonctionnement, une particularité rendant l'évaluation partielle. Les systèmes ne sont observés et examinés que dans une phase prématurée de leur fonctionnement qu'est l'extraction (de termes, de relations ou de classes). Les applications finales telles que l'indexation, l'enrichissement, la mémoire de traduction ou la veille... suivent dans une phase ultérieure qui nécessite d'autres connaissances, non mises à la disposition des organisateurs.

Cela explique que les organisateurs et les fournisseurs de systèmes sont bien conscients que l'évaluation serait partielle et ne porterait que sur une première phase du fonctionnement des systèmes. Une évaluation plus globale nécessiterait la mise à disposition des outils ; point non envisagé dans cette campagne. Malgré cette restriction, nous avons essayé de dégager trois tâches définies en termes *d'applications*.

L'appréciation (ou non) des performances d'un système ne peut être indépendante de l'application industrielle ou langagière pour laquelle le système a été conçu. L'application doit guider la conception même de l'outil ; c'est l'un des enseignements majeurs de la campagne Arc A3, et que nous avons toujours maintenu, ici dans le projet Cesart. Trois applications sont fixées (description ci-après) :

- T1 : Extraction de termes pour l'enrichissement de ressources terminologiques
- T2 : Indexation contrôlée et enrichissement
- T3 : Extraction de relations

2.3. Présentation des données textuelles

La campagne d'évaluation doit porter sur un corpus dont les propriétés physiques et thématiques (taille, format, structure, balisage, contenu, propriété d'homogénéité, …) répondent aux besoins et aux contraintes techniques et scientifiques des systèmes. Le corpus doit aussi être volumineux afin de satisfaire les systèmes fondés sur des calculs probabilistes, bien que seule une partie réduite fasse l'objet de l'expertise réelle des systèmes. Le corpus dans son intégralité sert alors au noyage. Des ressources complémentaires sont mises aussi à la disposition de la campagne selon les besoins des partenaires :

- Les outils fondés sur des statistiques reçoivent un corpus d'apprentissage similaire au corpus de l'évaluation. La similarité concerne la taille, le format et le thème.
- Les organisateurs disposent aussi d'un référentiel d'appariement. C'est une ressource validée par des experts, elle sert aux calculs des taux de rappel et de précision à partir des résultats des systèmes. C'est une liste terminologique en rapport avec le contenu du corpus, elle peut être structurée (comme un thésaurus) ou non (liste à plat). Elle n'est pas communiquée aux participants.

L'extraction de termes est souvent en amont par rapport aux fonctions des extracteurs de relations et des classifieurs. Autrement, l'acquisition de ressources terminologiques devrait être considérée comme une chaîne de production et de *coopération* entre lesdits outils (Béguin *et al.*, 1997). Dans une optique de mutualisation de ressources, on peut associer le processus d'extraction de relations ou de classes à celui de l'extraction de termes. Les deux procédés peuvent être interdépendants, dans la mesure où les classifieurs et les extracteurs de relations peuvent naturellement utiliser les résultats des extracteurs de termes et inversement, un extracteur de termes peut utiliser les liens (syntaxiques ou sémantiques) déterminés par les extracteurs de relations pour restructurer ses termes. Il ne s'agit donc pas de prendre la liste des termes séparés de leur(s) contexte(s) mais plutôt de garder la possibilité de retour au corpus pour réaliser l'extraction de relations ainsi que la constitution de classes. Certains logiciels dans leur état actuel ne permettent pas ce type de procédé[6]. Il faudrait envisager ce processus dans une chaîne de coopération si on voulait obtenir à partir du même corpus diverses ressources terminologiques.

- Pour la catégorie des extracteurs de relations, certains produisent eux-mêmes (dans une étape préliminaire) leur propre *liste de termes de repérage* qui servira à l'identification de relations[7]. D'autres ont besoin d'entrée de cette *liste de termes d'amorçage* en rapport avec le contenu thématique du corpus afin de s'y référer dans l'extraction des relations entre termes (exemple de Seek-Java (Le Priol, 2000)). Les organisateurs ont pris en considération ce besoin en ressource externe, lors du choix du corpus.

2.4. Présentation de l'expertise humaine

En concertation avec les participants, des experts humains sont désignés pour mener l'aspect manuel et qualitatif de l'évaluation. Vu l'aspect interdisciplinaire du projet CESART, une attention particulière sur les connaissances scientifiques et compétences pratiques des juges est à observer soigneusement. Le juge doit faire preuve d'une double compétence : spécialiste du domaine du corpus d'une part, et un familier des pratiques documentaires et langagières telles que l'indexation, la terminologie, la traduction… d'autre part.

[6] Il serait intéressant dans cette optique d'avoir des listes de termes avec ou sans lemmatisation. Cette possibilité (mise à disposition de l'utilisateur de la *forme de surface et de sa base*) est offerte par *IDE/XTS de Temis*.

[7] Cas de certains extracteurs de relations considérés également extracteurs de termes.

Chaque juge est invité à examiner un ensemble de résultats au maximum, ceci leur permettra d'éviter une surcharge mentale qui n'est pas sans incidence sur l'évaluation. Et pour que l'évaluation d'un système ne soit biaisée par la subjectivité d'un seul juge, il est envisagé de soumettre un même système aux regards de deux juges au minimum. Ce qui reste raisonnable pour pouvoir croiser les résultats des différents systèmes avec les différentes appréciations des évaluateurs.

Il est évident qu'une évaluation manuelle (jugement humain qualitatif) reste plus fiable dans la description des performances et des limites des systèmes, mais constitue toutefois un référentiel qui n'est malheureusement pas reproductible et ne permet pas d'évaluer le silence sauf si le protocole spécifie que les experts doivent relever les termes manquants non extraits par le système, tâche très fastidieuse.

À l'opposé, une évaluation automatique (basée sur un algorithme d'appariement à des référentiels externes), semble être *a priori* un procédé d'évaluation plus efficace que l'expertise humaine dans la mesure où, plus encore que l'expertise humaine et l'alignement linguistique, elle garantit la reproductibilité de l'expérience et par là même, l'obtention de résultats objectifs (Daille, 2002) et offre surtout une possibilité de procéder à des évaluations horizontales en raison de l'extensibilité du protocole.

De façon générale, l'expertise humaine et la constitution de ce matériel de test représentent un coût important dont il faut tenir compte lors du choix des corpus et de la définition des métriques dans le protocole.

De plus, l'évaluation humaine nécessite une interface adéquate. C'est un fait admis que lorsqu'un juge est placé devant une masse importante de termes ou de relations disposés à plat et dépourvus de leurs contextes, sans possibilité d'être renseigné sur d'autres points, il devient difficile de se prononcer avec exactitude sur la pertinence ou non des résultats fournis par les systèmes.

Dans le projet Cesart, nous mettons un dispositif d'interfaces pour la gestion *a posteriori* des résultats des systèmes. Ceci permet de dispenser les juges des lectures linéaires et fastidieuses des résultats, et procéder plutôt à des lectures en modes variés selon leurs besoins. Il s'agit d'une interface avec des options de visualisation :

- Navigation statique : le juge peut visualiser les termes par ordre lexicographique, fréquentiel, de pertinence ou par ordre de parution dans le texte. Dans le cas des extracteurs de relations, ce mode permet au juge de présenter entre autres les binômes extraits par typologie de relations.
- Consultation dynamique : le juge peut saisir des requêtes dans des champs en vue de se renseigner plus sur des termes ou des relations. Il peut demander le retour au contexte pour valider un terme ou chercher les relations liées à ce terme, il peut filtrer les données sur un champ plus restreint…
- Visualisation cartographique : une écriture/lecture graphique des données permet une visualisation plus claire des résultats favorisant la mémorisation des termes sous forme de champs lexical ou conceptuel et l'observation de la transitivité entre les relations.

Dans le cahier des charges soumis à l'appel Technolangue, nous avons prévu le développement de ce système. Mais il s'est avéré que parmi les participants, le système WorldTrek intègre un module d'export *Visuter* qui permet dans sa version actuelle de répondre à la majorité des points soulevés ci-dessus. Quelques ajustements restent toutefois à effectuer.

3. Calcul des métriques d'évaluation

3.1. Métriques pour la tâche de constitution de terminologie

Dans le cadre de Cesart, la tâche1 ou application essentielle vis-à-vis de laquelle les systèmes vont être évalués est la constitution de ***terminologie systématique***. Il s'agit de l'acquisition de termes et de relations en vue de la construction de ressources terminologiques (thésaurus).

Caractéristiques de la tâche

- Input (fichier HTML) : trois corpus relevant de trois domaines différents.
- Output (fichier XML, HTML, XLS) : liste ordonnée de termes contenant les renseignements suivants : forme canonique ; rang de pertinence ; variations ; fréquence ; lien pour retour au contexte ; puis toute information complémentaire qui pourrait être utile aux évaluateurs.
- Recommandation aux participants : donner les termes les plus importants susceptibles d'être de bons descripteurs pour élaborer un thésaurus d'indexation documentaire correspondant au corpus.

Recommandations aux juges

Pour l'évaluation manuelle, les experts reçoivent des résultats en papier et en version électronique. Pour rapporter leur jugement, ils disposent d'une grille de notation correspondant à trois niveaux dégressifs A, B et C de pertinence des termes. La note A indique que l'unité proposée est correcte et à retenir par rapport à l'application visée. La note B signale qu'il s'agit d'une unité candidate et n'est pertinente que sous conditions (par exemple le terme extrait doit subir une variation morphologique ou être mis ou retiré d'un syntagme l'englobant). Enfin la note C est pour signifier que le terme proposé est à rejeter car il ne correspond guère à l'application recherchée.

Une autre colonne est réservée pour les éventuels commentaires pouvant figurer devant une unité extraite. Il revient aux organisateurs de préciser le *taux de rappel* en calculant le rapport du nombre de termes correspondant à la note A par rapport au cardinal du référentiel d'appariement, puis le *taux de précision* en calculant le rapport du nombre de termes correspondant à la note A sur l'ensemble des termes extraits (A+C). L'inclusion (ou non) des termes correspondant à la note B (termes plus ou moins acceptables) dans le calcul des taux est assujetti à un traitement particulier.

En plus de cette évaluation quantitative, tout juge est invité à donner une classification *qualitative* de l'ensemble des systèmes expertisés en leur attribuant une note globale choisie entre 1 et 5 assortie d'une appréciation. Cet ordonnancement proposé par les experts permet de vérifier deux hypothèses :

- La subjectivité et le degré de tolérance des juges sont-ils sans incidence sur l'évaluation manuelle ? une différence importante dans l'ordonnancement global des systèmes implique une divergence dans les regards des juges et par suite dans l'appréciation et l'évaluation des systèmes de la campagne.
- L'appréciation globale donnée par les experts à l'issue de cet ordonnancement est-elle en concordance avec les mesures des taux de rappel et de précision calculées par les organisateurs ?

3.2. Métriques pour la tâche d'indexation contrôlée

L'évaluation se réalise en deux phases successives.

Caractéristiques de la tâche

- Input (fichier HTML) : 3 corpus relevant de trois domaines différents ; trois thésaurus d'indexation ad hoc ; codes abstraits (identifiants) pointant sur les concepts de chaque thésaurus, afin d'éviter les problèmes de comparaisons de chaînes et de normalisation terme/variantes.
- Output (fichier XML, HTML, XLS) : liste de termes référant à chaque texte du corpus, susceptibles de l'indexer, accompagnée des renseignements suivants : Identifiant ; variations ; fréquence ; retour au contexte ; puis toute information complémentaire qui pourrait être utile aux évaluateurs.

Phases d'évaluation

- Phase 1 : évaluation automatique par appariement avec le référentiel humain conçu pour cette même tâche ; calculer les taux de rappel/précision sur la liste des termes d'indexation attachés à chaque document dans le référentiel.
- Phase 2 : évaluation manuelle par un expert pour les termes d'indexation supplémentaires proposés par les systèmes, qui ne figurent pas dans le référentiel (termes attachés à chaque document). L'expert juge s'il est pertinent ou non d'inclure ces nouveaux termes.

Si de nouveaux termes sont déclarés pertinents par l'expert, alors on procède à une nouvelle évaluation automatique (comme dans la phase 1) sur le référentiel augmenté (enrichi).

Évaluation automatique

Si l'évaluation manuelle est considérée comme un dispositif valide et fiable dans les campagnes d'évaluation, la subjectivité des experts, leurs niveaux de connaissances et leurs degrés de tolérance sont souvent différents. Ceci constitue une des limites de ce mode et peut justifier le recours à une évaluation automatique comme complément. L'association des deux procédures permet en plus de vérifier la fiabilité et la crédibilité des métriques adoptées et mises en jeu dans l'opération, ce que nous pourrions appeler *l'évaluation de l'évaluation* (ou la *méta-évaluation*).

Dans l'évaluation manuelle susmentionnée, si le taux de précision dépend entièrement de l'appréciation de l'expert (notes A et C), le taux de rappel dépend conjointement de l'expert (note A), et d'un référentiel qui recense l'effectif des termes objets de silence. Dans l'évaluation automatique, les deux métriques sont inhérentes entièrement et uniquement au référentiel. Un programme informatique permet l'appariement des résultats des systèmes avec les termes du référentiel et en déduire systématiquement les taux de rappel et de précision.

Le progiciel d'appariement conçu dans le cadre Aupelf Arc A3 répond partiellement à cet objectif. Il est important d'élaborer une nouvelle version plus fiable surmontant les limites des comparaisons lexicographiques (de chaînes de caractères) en repensant aux appariements prenant en compte les variations morpho-syntaxiques des termes extraits.

3.3. Métriques pour la tâche d'extraction de relations

Si les métriques proposées ci-dessus sont implémentables sans difficulté majeure dans le cas des extracteurs de termes, la situation semble relativement complexe dans le cas des extracteurs de relations. Ces derniers sont difficilement comparables à cause de la diversité des types de relations extraites, la différence des modèles sémantiques implémentés et leur

influence sur les relations extraites[8], la différence des objectifs et des applications visés. Le protocole fondé sur les mesures de performance dont nous avons fait l'expérience lors de notre précédent projet nous impose de réfléchir à quelques adaptations :

- On peut distinguer le *bruit* (et par conséquent la précision) à condition qu'on ait une bonne connaissance des relations morpho-syntaxiques et sémantiques possibles du domaine. Une expertise humaine combinant la connaissance du domaine ainsi que des modèles sémantiques devient une nécessité. Ceci semble réalisable dans le contexte du corpus que nous avons choisi.

L'évaluation de cette tâche se réalise en deux phases successives.

Caractéristiques de la tâche

- Input (fichier HTML) : 3 corpus relevant de trois domaines différents ; trois listes de termes d'amorçage correspondant à chaque domaine ; sorties fournies par Lexter et par Termic sur les documents des corpus ; étiquettes morphosyntaxiques sur les corpus.
- Output (fichier XML, HTML, XLS) : triplet [terme source, RELATION, terme cible] ; retour au contexte ; occurrences ; puis toute information complémentaire qui pourrait être utile aux évaluateurs.
- Recommandations aux participants : restreindre l'évaluation aux trois types de relations : synonymie, hyperonymie, méronymie, puis retrouver tous les triplets possibles dont l'un des termes appartient à la liste d'amorçage, donnée en entrée.
- Évaluation manuelle : l'expert donne une note sur une échelle à 3 valeurs pour mesurer la pertinence de la relation.

4. Discussion

Certes, nous insistons dans le cadre de Cesart sur l'ergonomie linguistique et informationnelle soutenue par plusieurs chercheurs. Une ergonomie qui se définit à travers un certain nombre de paramètres à prendre en compte, en particulier les classes d'applications (Chaudiron, 2001)[9]. De ce fait, la dimension d'usage[10] s'impose et elle parait particulièrement pertinente dans le contexte d'acquisition de ressources terminologiques. On se demande alors quel est le statut de juge responsable de la validation, doit-il être l'utilisateur final ou le spécialiste du domaine ?

D'autres tâches des systèmes d'acquisition de ressources terminologiques restent à étudier et soulèvent quelques questionnements :

Mémoire de traduction : Comment évaluer les outils de construction de lexiques terminologiques multilingues ? Sur la base d'un texte source traduit et aligné avec son texte cible correspondant, dans quelles mesures la traduction des termes extraits par des systèmes à

[8]Nous faisons abstraction sur les sources de cette divergence, étant donné la nature et le mode d'évaluation (boîte noire), qui ne tient pas compte des approches et des modèles adoptés par les systèmes évalués.

[9]L'auteur met l'accent sur les rôle des usagers et son impact sur l'évolution du paradigme de l'évaluation : « Si l'évaluation des logiciels correspond à une pratique codifiée et parfois normalisée d'analyse d'un objet, la multiplication des attentes des usagers d'une part et les différents courants de recherche issus de la plupart du mouvement «social informatics» américain permettent de considérer la pratique évaluative comme le point de départ d'un travail de réflexion sur l'acceptabilité des outils techniques issus des NTIC et, plus généralement, sur les rapports de l'homme à la machine. (sic).

[10]L'appel à propositions de Technolangue dans lequel s'inscrit le projet Cesart, a mis l'accent sur l'importance et la pertinence de la dimension de l'usage pour les campagnes d'évaluation.

partir du texte source peuvent-ils être considérés comme termes pertinents dans la langue cible ? Idem pour les extracteurs de relations.

La reformulation en recherche d'information : comment évaluer les extracteurs de relations dans l'optique d'une activité de recherche d'informations ? Dans quelles mesures des termes mis en relation par extraction automatique peuvent-ils servir de termes d'expansion (lexico-syntaxique ou sémantique) en vue d'améliorer la pertinence d'un système de recherche d'information par reformulation paraphrastique de requêtes (Timimi, 1998). Il s'agit d'une tâche qui nous semble intéressante mais non encore expérimentée : dans une démarche de recherche d'information, on peut par exemple proposer en entrée des termes (supposé appartenir à des requêtes d'utilisateurs) et demander aux système d'extraire à partir du corpus des termes proches (en relation) des termes des requêtes dans une optique de reformulation de requêtes ou dans une optique d'aide aux enrichissements de thésaurus.

Fouille de texte pour des applications de veille scientifique et technologique **:** Dans quelles mesures et de quel genre de référentiel doit-on disposer pour évaluer des systèmes basés sur des connaissances terminologiques et destinés à des fins de veille ?

Remerciements

Les auteurs remercient Pierre Zweigenbaum (STIM/DSI/AP-HP) pour sa participation dans la réflexion sur les métriques d'évaluation dans la cadre du projet CESART.

Références

Amar M., Béguin A., David S. Debrito M., L'Homme M.-C., Mustafa El Hadi W., Paroubek P. and Timimi I. (2001). *Évaluation d'outils de construction automatique de termes et de relations sémantiques à partir de corpus en français*. Rapport scientifique du projet AUF ARC A3, ex-Aupelf-UREF, avril 2001.

Béguin A., Jouis C. and Mustafa El Hadi W. (1997). Évaluation d'outils d'aide à la construction de terminologie et de relations sémantiques entre termes à partir de corpus. In *JST'97*, FRANCIL, AUPELF-UREF, Avignon, avril 1997 : 419-426. Cet article a été publié dans Chibout *et al*. 2001 (eds.) : 161-179.

Bourigault D. and Jacquemin C. (2000). Construction de ressources terminologiques. In *Ingénierie des langues*, Traité 1C2 - Section informatique et systèmes d'information, Sous la dir. de Pierrel, J.-M, 2000 : 215-233.

Cavazza M. (1993). *Méthodes d'évaluation des logiciels incorporant des technologies d'informatique linguistique*. Paris, Rapport MRE-DIST, 1993.

Chaudiron S. (2001). *L'évaluation des systèmes de traitement de l'information textuelle : vers un changement de paradigmes*, Mémoire pour l'habilitation à diriger des recherches en sciences de l'information, Université de Paris 10, novembre 2001.

Daille B. (2002). *Découvertes linguistiques en corpus*. Mémoire d'habilitation à diriger des recherches en informatique, Université de Nantes.

Grefenstette G. (1994). *Explorations in Automatic Thesaurus Discovery*. Boston, Kluwer Academic Press, 320 p.

Le Priol F. (2000). *Extraction et capitalisation automatiques de connaissances à partir de documents textuels. SEEK-JAVA : identification et interprétation de relations entre concepts*, Paris, Thèse de doctorat, Université Paris-Sorbonne.

Mustafa El Hadi W., Timimi, I. Béguin A. and Debrito. M. (2001). The ARC A Project : Terminology Acquisition Tools : Evaluation Method and Tasks. In *Evaluation Methodologies for Language ad Dialogue Systems Workshop*, ACL/EACL, Toulouse, 6-7 Juillet 2001 : 41-50.

Sparck-Jones K. and Gallier J.R. (1996). *Evaluating Natural Language Processing Systems : An Analysis and Review*. Springer, Berlin, 1996.

Timimi I. (1998). 3AD : un outil de classification à caractère linguistico-mathématique. *Colloque annuel de Veille Stratégique Scientifique et Technologique* (*VSST'98*), IRIT et Delta Veille dans FAUST (Forum des Arts de l'Univers Scientifique et technologique), Toulouse, 19-23 Octobre 1998 : 299-310.