Revue COSSI

N° 9 | 2020
Les enjeux et les défis de la fonction d'évaluation en sciences de l'information et de la communication

Évaluation d'outils et outils d'évaluation

Les paradigmes d'évaluation, à l'épreuve …–

Ismaïl Timimi
Maître de Conférences en SIC
(Sciences de l'Information et de la Communication)

Laboratoire Geriico, Université de Lille

Édition électronique :
URL : https://revue-cossi.numerev.com/articles/revue-9/1789-evaluation-d-outils-et-outils-d-evaluation
DOI : 10.34745/numerev_1560
ISSN : 2495-5906
Date de publication : 29/11/2020

Pour **citer cette publication** : Timimi, I. (2020). Évaluation d'outils et outils d'évaluation. *Revue COSSI*, (9). https://doi.org/https://doi.org/10.34745/numerev_1560

Parce qu'elle peut servir la planification ou l'aide au développement, la certification ou la mise en conformité, la gratification ou la promotion... l'évaluation a toujours suscité un intérêt particulier de la part des professionnels et des chercheurs dans différents domaines et disciplines scientifiques. Cet article s'intéresse aux paradigmes de l'évaluation dans le cas des Sciences de l'Information et de la Communication, et particulièrement aux systèmes de traitement automatique de l'information, évalués à des fins info-documentaires. Dans une première partie, nous présentons une réflexion sur les pratiques de l'*évaluation* en tant que méthodologie de recueil et d'analyse de données et les problématiques sous-jacentes. Pour concrétiser nos propositions et justifier nos réflexions, nous nous appuyons sur une revue de littérature recensant quelques programmes et actions d'évaluation organisées aux niveaux national et international. Une seconde partie est consacrée aux discussions et débats autour de l'*évaluation* en tant que préoccupation scientifique entre un souci de théorisation et une complexité de normalisation.

Mots-clefs :
Pertinence, Métrologie, Traitement d'information, Pratiques et usages, Référentiel, Paradigmes d'évaluation

---

Abstract : Evaluation has always attracted special interest among a large variety of professional groups, scientific disciplines and other related circles. It is because it can be used for planning or assistance development, certification or compliance, gratification or promotion it has gained this momentum. This article focuses on the evaluation paradigms within Information and Communication Sciences, and in particular automatic information processing systems, evaluated for info-documentary purposes. In the first section, we present a state of the art of evaluation practices and define a methodology for data collection and analysis together with more related and fundamental issues. In order to concretize our reflections and proposals, we rely on a literature review listing some evaluation campaigns and actions organized at the national and international levels. The second part of our article will be devoted to discussions and debates on evaluation as a scientific concern between a concern about theorization and the complexity of standardization.

Keywords: metrology, evaluation paradigms, information processing, relevance, practices and uses, reference system.

INTRODUCTION

Parce qu'elle peut servir la planification ou l'aide au développement, la certification ou la mise en conformité, la gratification ou la promotion... l'évaluation a toujours suscité un intérêt majeur de la part des industriels et des chercheurs dans différents domaines professionnels et disciplines scientifiques.

Selon des contextes, sa dénomination est assimilée parfois à ses finalités (*valorisation, validation, notation, labélisation, qualimétrie, webométrie...*) et sa modélisation s'est construite autour de supports aussi variés (*audit, benchmark, tableau de bord, matrice swot, diagramme de Gantt...*) (Reider, 2000).

Ses concepts et paramètres sont associés à un réseau sémantique étendu (*performance, pertinence, distance, référentiel, métrique, besoin informationnel*) générant ainsi des dichotomies et une querelle de méthodes en opposition (*quantitative/qualitative, automatique/manuelle, verticale/horizontale, boîte transparente/boîte noire, ex-ante/ex-post, intrants/extrants, interface dynamique/ à interface statique, orientée système/orientée usager*...).

Quand un système est supposé à partir de son traitement d'information servir à des applications et des usages préconisés, il devient indispensable de vérifier par une évaluation la légitimité et l'exactitude de cette préconisation. L'évaluation devient alors comme une procédure pour juger un ou plusieurs attributs : l'adéquation du système à l'usage préconisé, la conformité de ses résultats par rapports aux attentes, son taux de progression vers le but déterminé ... et manifestement, c'est dans la modélisation de cette procédure et dans son cadrage méthodologique où réside toute la difficulté de l'évaluation aussi bien sur le plan théorique que pratique.

Si l'évaluation en tant que *pratique* est massivement exercée dans le milieu universitaire, en tant qu'*objet d'étude* elle n'a pas eu le même mérite. En sciences de l'information et de la communication (SIC), peu nombreux sont les travaux de recherche qui se sont intéressés à l'objet de l'évaluation des systèmes d'information et à fortiori des systèmes de traitement automatique de l'information. Paradoxalement, on constate dans l'histoire des sciences de l'information que la recherche documentaire est considérée comme un des domaines pionniers où se sont développées les premières réflexions atour de l'évaluation quantitative, des réflexions qui ont été à l'origine des métriques traditionnelles (Rappel et Précision) et de leur moyenne harmonique F-measure (Van Rijsbergen, 1979 ; Salton et McGill, 1983). Aujourd'hui la volumétrie a évolué, les besoins se sont diversifiés et les problématiques se sont complexifiées, mais l'on conserve encore les mêmes indicateurs de référence depuis des décennies.

Dans cet article centré sur l'évaluation en tant qu'objet d'étude, nous essayons dans une première section de recadrer ce concept à travers ses différents modèles et attributs, et particulièrement dans les contextes info-documentaires. Ensuite, dans une

deuxième section considérée comme terrain expérimental, nous présentons le fonctionnement et l'apport d'un ensemble de systèmes de traitement d'information (à base linguistique) en présentant certains programmes et actions d'évaluation, qui leur ont été consacrées au niveau national et international. Enfin, une dernière section est consacrée aux discussions et débats autour de l'évaluation et son cadrage méthodologique.

L'ÉVALUATION… DES CONCEPTS ET DES MÉTHODES

Les TIC et l'évaluation dans les SIC

Dans une perspective d'analyse et d'expertise, et contrairement aux méthodologies classiques de recueil et d'analyse de données telles que l'analyse de traces, l'entretien, le questionnaire, l'observation…, l'évaluation dans le cas des systèmes de traitement d'information présente un travail de recherche multidisciplinaire assez préoccupant. Le manque d'une normalisation consensuelle additionné à la difficulté d'une modélisation rend l'évaluation comme objet de recherche complexe et très discutable dans les différents champs disciplinaires, tels que les sciences économiques et sociales, sciences de gestion, sciences politiques, sciences de l'éducation, sciences du langage, informatique et mathématiques et davantage dans les sciences de l'information et de la communication...

Située au carrefour de ces différents champs professionnels et disciplinaires, l'évaluation est souvent assimilée à ses finalités, elle peut avoir le sens de *valorisation* dans les activités économiques et marketing ; le sens de *certification* dans des activités d'assurance qualité ; le sens de *notation*, en DRH ou plus généralement pour tout domaine ayant recours à une échelle ou un référentiel de cotation. Ces concepts et d'autres sont assez récurrents dans les travaux de recherche sur l'évaluation, aussi bien sur le plan méthodologique qu'empirique.

En sciences de gestion, on a recourt à une évaluation multicritère ACV (*analyse du cycle de vie*) qui a pour ambition de quantifier l'ensemble des impacts d'activités humaines sur l'environnement en déterminant de manière systématique les consommations de ressources et les émissions de substances liées à la production d'un bien ou d'un service (Guérin-Schneider et Tsanga Tabi, 2017).

Dans le domaine des ressources humaines, l'évaluation souvent individualisée vise à mesurer les compétences et performances du personnel dans le cadre d'une gestion de carrières professionnelles ou d'un plan de formation. Le cadre méthodologique s'appuie généralement sur les méthodes et techniques couramment utilisées en sciences économiques et sociales (questionnaire, entretien, observation…)

En sciences de l'éducation, l'évaluation qu'elle soit diagnostique, sommative ou formative (Michel et Rouissi, 2003), a toujours eu une place prédominante, notamment sous l'effet de la prégnance du digital et des innovations pédagogiques. Dans leurs travaux de recherche sur les pratiques de l'évaluation et aussi sur l'évolution de la

recherche en évaluation, Jorro et Droyer (2019) se sont intéressés aux formes évaluatives en usage avec ses enjeux nouveaux et ses obstacles récurrents.

Si les TIC en tant qu'*objets d'étude* ont suscité constamment un intérêt primordial en SIC, les chercheurs en sciences de l'information et du document, et encore moins en sciences de la communication, ont très peu exploré l'évaluation de ces technologies dans leurs préoccupations scientifiques. Au mieux, ils la survolent brièvement, quand ils s'intéressent dans leurs travaux de recherche à l'analyse des usages, des pratiques, des dispositifs, des processus de ces TIC.

Dans une logique interdisciplinaire à laquelle notre discipline SIC s'ouvre par tradition, il nous a semblé pertinent de considérer les travaux réalisés sur l'évaluation par des disciplines connexes et d'en croiser les approches et modèles. Si l'évaluation des systèmes en sciences expérimentales a suscité des intérêts scientifiques dirigés par les soucis de la conception et du développement, en SIC il a fallu réinterroger l'évaluation en lui donnant d'autres dimensions scientifiques, tant sur le plan méthodologique en s'intéressant à la nature et à la pertinence des critères d'évaluation et aux dispositifs mis en œuvre que sur le plan théorique et épistémologiques en réinterrogeant les notions de métrique, de pertinence, de satisfaction d'un besoin informationnel et le dilemme entre l'évaluateur usager et l'usager évaluateur.

Les facettes de l'évaluation

Généralement, l'évaluation d'un système est l'appréciation de ses performances sur la base d'un besoin informationnel à satisfaire, à partir de ressources intrinsèques ou extrinsèques mises à disposition... la diversité des besoins et des ressources est à l'origine d'une typologie des formes d'évaluation.

Dans une évaluation dite de *progression* (*verticale*), un 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, et inscrite dans les études de génie informatique... À l'opposé, une évaluation dite d'*appariement* (*transversale*) consiste à comparer les performances d'un système par rapport à d'autres conçus pour des applications similaires. Cette démarche est très courante dans les études de benchmarking et de veille concurrentielle et technologique. Il se peut aussi que cet appariement transversal ne soit pas inter-systèmes mais effectué plutôt par rapport à des référentiels prédéfinis, établis manuellement ou autrement, mais surtout validés.

Dans une autre optique d'évaluation, dite de *diagnostic* (*boîte transparente* ou *glass box*), l'usager cherche à déterminer à partir d'une série de tests les sources de performance ou d'imperfection d'un système par rapport à une tâche précise. Dans cette démarche, minutieuse et complexe, l'évaluation ne concerne pas le système dans sa globalité mais seulement certains modules (de prétraitements) intrinsèques évalués parfois séparément. Cette démarche, elle aussi, est orientée conception et développement dans la mesure où ces tests de diagnostic permettent de réviser et

développer par progression les performances d'un système à partir d'une évaluation de ses composantes. À l'opposé, une évaluation peut être menée sur le concept de la *boîte noire* (*black box*). Elle consiste à faire abstraction sur les composantes intrinsèques du système, et ne s'intéresse qu'au jugement des performances globales du système, elle se focalise uniquement sur les ressources mises en entrée du système (*Input*) et sur les données obtenues en sortie (*Output*). Les prétraitements des données effectués par les différents modules du système ne font l'objet d'aucune évaluation dans cette démarche. Pour des raisons expliquées ultérieurement, l'évaluation selon le principe de la *boîte noire* est celle adoptée unanimement par la plupart des campagnes d'évaluation.

Concernant l'interactivité de l'usager, les études d'évaluation lui réservent une place considérable. L'évaluation à *interface statique* consiste à juger les performances d'un système, sans faire appel à des interventions humaines ou à des enrichissements extérieurs. À l'inverse, l'évaluation à *interface dynamique* permet d'étudier la valeur ajoutée et l'impact des ressources d'enrichissement introduites dans le système (ex. bases de connaissances ou mémoires de traduction) ou des choix d'orientation ordonnés par l'usager (ex. à des fins d'apprentissage automatique).

Une évaluation peut être *quantitative*, exprimée par des *métriques* qui calculent le degré de similarité entre les données fournies en sortie par le système et les référentiels préétablis. À l'opposé, une évaluation *qualitative* consiste à analyser et annoter les performances des systèmes sans forcément les exprimer en notation numérique.

Parallèlement à cette dernière dichotomie, on trouve aussi l'évaluation *manuelle* versus *automatique* selon les usages. Si l'expertise manuelle est considérée comme un dispositif valide et fiable dans les actions et programmes d’évaluation, la subjectivité des experts, leur niveau de connaissances et de pratiques, et leur degré de tolérance sont souvent sources de questionnements et de débats dans les paradigmes d'évaluation. En revanche, la volumétrie des résultats fournis par les systèmes pour certaines applications contraint les experts à recourir parfois à une évaluation automatique, outillée par un algorithme d'appariement des données fournies par le système à des référentiels préétablis et stables. Les limites ou plutôt la complémentarité des deux approches peuvent justifier le recours à une solution hybride, qui permettrait en plus de vérifier la fiabilité et la crédibilité des métriques adoptées et mises en œuvre dans le protocole (Timimi, 2006).

Évaluation et tendances des protocoles

Face à ces dichotomies de processus et d'outils, l'évaluation reste une fonction modulable dans la mesure où les méthodes ne sont pas forcément cloisonnées ou exclusives mais peuvent être combinées et adaptées selon les enjeux et les contextes des actions d'évaluation d'une part et la typologie et contraintes des systèmes participant d'autre part. Dans la majorité des protocoles étudiés, on relève des évaluations multimodales (ex. *horizontale, à interface dynamique, quantitative et selon le principe de la boîte noire*…).

Généralement, dans la plupart des protocoles d'évaluation étudiés, on constate que l'évaluation *verticale* (*de progression*) est abandonnée, dans la mesure où il est difficile aux organisateurs de disposer des systèmes et encore moins de leurs versions antérieures pour pouvoir étudier l'évolution diachronique des performances. L'évaluation sur le principe de la *boîte transparente* (de *diagnostic*), elle aussi, est moins utilisée car il s'agit d'un dispositif difficile à mettre en place, il requiert une connaissance des processus internes et des fondements théoriques de chacun des systèmes participant. Il réclame l'accès à l'architecture et au code du développement du système, ce qui risque d'être compromettant lorsque l'évaluateur est un intervenant extérieur (Cavazza, 1993). Cette distinction semble justifiée vu que ces méthodes abandonnées sont orientées "conception" ce qui n'est pas l'objectif principal des campagnes d'évaluation et relève plutôt des services de développement et de maintenance propres à chaque système.

À l'opposé, le principe de l'évaluation *boîte noire* reste une pratique très courante dans la plupart des protocoles d'évaluation étudiés. Ce choix est justifié du fait qu'il s'agit d'un dispositif d'expertise facile à mettre en œuvre, unanimement accepté dans un consortium composé de chercheurs universitaires, d'industriels de systèmes et d'usagers potentiels, et pose le moins de problèmes méthodologiques et d'obligations empiriques. Sans nécessiter l'accès au fonctionnement interne des systèmes, ce choix permet une étude comparative malgré la différence des architectures et des prétraitements employés. (Cavazza, 1993 ; Sparck-Jones et Gallier, 1996).

LES CAMPAGNES D'EVALUATION COMME TERRAIN D'OBSERVATION

Campagnes d'évaluation : État des lieux

Peu connue comme activité normalisée, l'évaluation des systèmes de traitement automatique d'information est restée au centre des préoccupations d'organisations institutionnelles, politiques, industrielles et scientifiques. Plusieurs programmes et actions d'évaluation ont été organisés au niveau national et international et donné lieu à des publications et congrès dédiés exclusivement à l'évaluation (*EACL, ACL, LREC*...). La plupart de ces actions se sont intéressées principalement aux systèmes de traitement automatique de l'information textuelle (écrite ou orale).

Le traitement de l'information textuelle, comme d'ailleurs celui de l'information imagée, compte un grand nombre de systèmes, conçus majoritairement sur des bases linguistiques et/ou statistiques et sous une architecture multi-agents SMA (*système multi-agents*). À partir d'une modélisation des processus langagiers, les systèmes sont développés dans une perspective d'automatiser des pratiques et des activités informationnelles et documentaires, prises en charge habituellement par des usagers (humains). En fonction des besoins informationnels et de la complexité de modélisation, cette automatisation peut être partielle ou totale.

La dichotomie entre un traitement automatique et un traitement manuel a été alors et demeure le stimulateur principal pour la plupart des programmes d'évaluation au niveau international comme CLEF (*Cross Language Evaluation Forum*) ; DARPA (*Defense Advanced Research Projects Agency*) ; MUC (*Message Understanding Conferences*) ; NTCIR (*NII Test Collection for IR Systems*) ; TDT (*Topic Detection and Tracking*) ; TREC (*Text REtrieval Conference*), et dans une moindre mesure, au niveau national et francophone, comme le *Programme TECHNOLANGUE* (Paroubek et al., 2007).

Retour sur les composantes d'un outil à base linguistique

Généralement, un système de traitement automatique d'information s'appuie d'abord sur un prétraitement technique (formatage, nettoyage...), ensuite sur un prétraitement linguistique faisant appel à un ensemble de modules :

- un module de segmentation du corpus en unités d'analyse (en paragraphes, en phrases, en propositions autour d'un verbe, en syntagmes, en concepts, en mots...)
- un module d'annotation morphosyntaxique qui consiste à analyser chaque forme du corpus dans son contexte et lui associer son étiquette morphosyntaxique ; cela permet de déceler des relations potentielles entre des mots du texte ayant des orthographes différentes (ex. reines, régner, royal) et surtout d'anéantir des relations entre des mots du texte ayant des orthographes identiques (ex. *résident=verbe, résident=adjectif, résident=nom*) ou (ex. *poste=verbe, poste=nom.mas, poste=nom.fém*).
- un module d'analyse sémantique qui consiste à identifier des relations de synonymie entre les constituants du texte, voire extraire des réseaux sémantiques.

En fonction de l'architecture du système, ces modules interviennent de manière séquentielle ou interactive pour lever les ambiguïtés qui marquent profondément les langues naturelles, des mots de même orthographe mais de sens différents (polysémie) et des mots de même sens mais d'orthographes différentes (synonymie), en plus des autres difficultés liées aux anaphores, métaphores....

Les paradigmes d'évaluation et les travaux empiriques des campagnes

Au-delà de la différence des modèles théoriques et architectures, nous pouvons étudier l'évaluation de systèmes sous trois grandes entrées, inspirées de la norme ISO 9126 (ISO, 1991). On peut évaluer les paramètres internes du système, sans nécessairement l'exécuter ; on peut ainsi étudier ses dictionnaires et grammaires, ses algorithmes, le volume et la nature des données à traiter, etc. Il reste toutefois difficile d'estimer l'impact et la valeur ajoutée de ces paramètres. D'ailleurs, c'est pour cette raison que la plupart des campagnes d'évaluation utilisent plutôt les paramètres externes et proposent dans leur protocole des approches pour étudier la fonctionnalité, la fiabilité, l'utilisabilité, l'efficacité, la maintenance ou la portabilité d'un système. Enfin, on peut

évaluer un système en fonction de contexte pour s'intéresser à des paramètres liés davantage à l'usage tels que l'efficacité, l'efficience (rendement), la satisfaction, ou la sureté (Popescu-Belis, 2007).

De même, nous pouvons étudier l'évaluation de systèmes en deux autres entrées, qui sont cette fois-ci davantage liées aux types de besoins auxquels répondent ces systèmes. a) Il s'agit de systèmes linguistiques, proprement dits, comme des outils de désambiguïsation lexicale ou des analyseurs morpho-syntaxiques (tagging, parsing). Ces systèmes ne présentent aucun intérêt pour un usager final, si ce n'est qu'à des concepteurs ou des chercheurs en ingénierie linguistique. Leur évaluation n'a d'ailleurs pas suscité de grands intérêts malgré le succès de la campagne d'évaluation GRACE (*Grammaires et Ressources pour les Analyseurs de Corpus et leur Evaluation*) lancée vers les années 1994, par le CNRS. Si cette action n'a jamais été reconduite, elle a le mérite d'être considérée comme projet précurseur de toutes les campagnes d'évaluation effectuées ultérieurement au niveau national et qui a permis de fonder les premières réflexions sur l'étude de l'évaluation tant sur le plan méthodologique que scientifique. b) Il s'agit des outils applicatifs : ce sont des outils basés certes sur des prétraitements linguistiques et fournissent en sortie des données directement exploitables à des fins info-documentaires tels que des outils de catégorisation ou classification, de filtrage, d'indexation, d'extraction de termes d'un domaine. On compte ici également des systèmes avec des applications telles que l'*Analyse de sentiments ou d'opinions, la recherche d'information, le résumé et la traduction automatique...*

Concrètement, la plupart des campagnes d'évaluation observées ont adopté un paradigme d'évaluation fondé sur un processus en deux phases : d'abord une mise à disposition des données nécessaires aux systèmes ensuite une restitution des résultats fournis par les systèmes en réponse à une même tâche (dite *de contrôle)* communiquée en amont comme une hypothèse d'usage. Toutes les évaluations sont partielles et ne portent que sur des tâches de contrôle (une partie des applications des systèmes). Une évaluation plus globale serait onéreuse et difficile tant sur le plan logistique que scientifique, son coût peut être excessivement plus élevé que le bénéfice retiré. Le présupposé du paradigme est qu'il est possible de mesurer en contraste l'efficacité des systèmes dans un domaine précis en définissant une tâche de contrôle à la fois proche des applications potentielles, pour des retombées industrielles, mais suffisamment générique pour convenir à la majorité des acteurs industriels et académiques de la campagne. Aussi, l'hypothèse sous-jacente dans la plupart des paradigmes d'évaluation d'ici, est que dans le cas d'une tâche de contrôle suffisamment représentative d'une problématique et d'un besoin informationnel réel, une différence quantitative significative entre deux systèmes traduit nécessairement une différence qualitative entre leurs modèles et bases théoriques (Adda et al., 1999).

Ainsi les systèmes de traduction automatique (dans la campagne CESTA) ont été évalués sur des tâches comme la lisibilité et grammaticalité du texte produit, la fidélité sémantique ; les systèmes questions-réponses (dans la campagne EQUER), ont été évalués sur des tâches comme les questions factuelles, les réponses binaires, les requêtes de définition ou d'énumération. D'ailleurs, c'est le changement récurrent de

tâches et la définition de nouvelles tâches traduisant de nouveaux besoins, qui fait l'objet et l'argument principal de chaque nouvelle édition d'une campagne.

LA MÉTA-ÉVALUATION… OU L'ÉVALUATION À L'ÉPREUVE

Dans cette partie, nous présentons et discutons les différents paramètres des cadres méthodologiques d'évaluation, le corpus et l'échantillon comme ressources textuelles d'entrée, puis les référentiels et les experts comme repères de comparaison et de jugement. Nous réservons une discussion particulière aux métriques employées dans les projets.

Le corpus, un matériel souvent conditionné

Une réflexion sur la constitution de corpus spécialisés est une phase indispensable dans tout projet d'évaluation d'outils de traitement d'information. D'après Pincemin (1999), le corpus doit vérifier trois types de conditions : *signifiance, acceptabilité et exploitabilité* en plus de la pertinence par rapport à un objectif d'analyse. L'ensemble de ces conditions est nécessaire pour sa réutilisabilité.

Dans un programme d'évaluation, les organisateurs mettent généralement 3 types de corpus à disposition des systèmes : un corpus d'entraînement (dit aussi de *test à blanc* ou de *training phase*) qui permet une préparation des systèmes à partir d'une simulation des paramètres technique et logistique de l'évaluation officielle ; un corpus de masquage qui permet d'accroître le volume des données à traiter et de dissimuler ainsi la partie dédiée à l'évaluation ; le corpus d'évaluation qui doit avoir des propriétés physiques et thématiques (taille, format, structure, balisage, contenu, homogénéité, …) répondant suffisamment aux contraintes techniques et scientifiques des systèmes. Des ressources complémentaires peuvent être aussi mises à disposition de la campagne selon les besoins des systèmes comme des données d'apprentissage pour les systèmes à base statistique (*dry run*), ou des termes d'amorçage pour les systèmes d'extraction de relations (Le Priol, 2000).

L'échantillonnage, un outil parfois fiable, parfois contestable

Dans la plupart des programmes d'évaluation basés sur une expertise humaine, il n'est pas évident que les évaluateurs effectuent un travail d'appréciation sur tout l'ensemble des résultats rendus par les systèmes. Et aussi pour des raisons de surcharges cognitives qui conduisent parfois à des hésitations répétées et finalement à des jugements arbitraires, il est fortement recommandé de procéder par échantillonnage, dans le sens où l'échantillon à évaluer peut être choisi à partir des données en entrée, ou plutôt à partir des résultats de sortie.

Seulement, cela n'est pas sans interrogations : que l'échantillon soit sélectionné dans les données d'entrée ou dans les résultats de sortie, comment justifier le choix de sa

taille et quels critères retenir pour garantir sa représentativité et déterminer son intervalle de confiance ?

Dans (Pincemin, 1999), cette règle de représentativité est bien commentée : « *on peut, lorsque le matériel s'y prête, effectuer l'analyse sur échantillon. L'échantillonnage est dit rigoureux si l'échantillon est une partie représentative de l'univers de départ* ». Dans ce cas, les résultats obtenus sur échantillon seront généralisables à l'ensemble de l'univers. Cependant, peu de travaux de recherche en SHS ont traité cette question et discuté les intervalles de confiance.

Pour atténuer cette limite, certains projets de recherche se sont appuyés sur des automates de validation automatique, qui permettent de comparer les données issues des systèmes à des référentiels préétablis. Certes, cette procédure a le défaut de réduire l'appréciation à un jugement binaire, mais elle a le mérite de pouvoir traiter l'intégralité des résultats donnés par les systèmes (et non pas seulement un échantillon) et reste toutefois un indicateur sur le comportement des systèmes face à l'ensemble du corpus. Dans d'autres contextes, ces automates de validation automatique permettent également d'approcher la valeur du rappel (taux des réponses pertinentes non identifiées par le système), une mesure impossible à calculer dans le cas d'une validation manuelle.

L'évaluation par référentiels, un débat toujours non achevé

Dans un programme d'évaluation, le choix du domaine et des thématiques ne peut être arbitraire. Il faut s'assurer de la disponibilité d'une part des ressources d'entrée adéquates et des référentiels validés dans le domaine sélectionné (pour une évaluation quantitative) et d'autre part des usagers évaluateurs familiarisés avec le domaine sélectionné (pour une évaluation qualitative).

Si le recours à des référentiels préétablis est la démarche la plus plébiscitée et la plus utilisée pour développer un cadre méthodologique acceptable dans un programme d'évaluation, un questionnement sur le statut de ces référentiels s'impose.

En étudiant une des campagnes d'évaluation (CESART) basée sur le principe des référentiels, on a constaté quelques dilemmes. Un système peut être proche d'un référentiel plus que d'un autre même si les deux référentiels sont valides et établis pour un même objectif. De même, un système peut être jugé proche des référentiels d'un domaine et non de ceux d'un autre. Une des solutions consiste à diversifier les référentiels jusqu'à l'obtention d'une saturation et d'une stabilité des résultats de systèmes, mais cette solution reste très coûteuse pour sa mise en œuvre. Par exemple, lors de l'évaluation de systèmes de construction de ressources terminologiques, les organisateurs ont du faire face à des référentiels institutionnels extrêmement différents, qui recouvrent cependant un même domaine qu'est l'éducation, ceci est du probablement aux "*variabilités des pratiques*" dans la construction et validation des référentiels.

Toutefois, la notion de "référentiel" reste encore en soi problématique. Est-il suffisant de procéder par comparaison de résultats donnés par des systèmes automatiques à des référentiels élaborés par des experts humains, pour en déduire de la qualité des systèmes ? Cette forme de comparaison n'est-elle pas réductrice dans la mesure où les résultats des outils, souvent conçus dans un but d'assistance, sont ici injustement jugés face à la qualité pertinente d'un travail humain validé, particulièrement dans le cas des pratiques professionnelles en information et documentation.

Si la comparaison à un référentiel humain peut paraître contestable dans la mesure où les systèmes sont souvent mal classés derrière les listes de référence, un grand nombre de campagnes ont privilégié de procéder autrement et de comparer les systèmes uniquement entre eux sans aucun référentiel extrinsèque. Cette évaluation inter-systèmes permet de créer un référentiel de consensus à partir des résultats communs fournis par la majorité des systèmes (*vote majoritaire* ou *pooling method*), cette solution convient également à défaut d'un référentiel de notoriété valide. Seulement, ledit référentiel commun post-édité, lui non plus, n'est pas sans limite, dans la mesure où il désavantage systématiquement les systèmes "*hors commun*". Tout système qui, lui seul, donne des résultats pertinents sera sanctionné s'il est comparé à un référentiel construit par un vote majoritaire. Ce point est resté problématique dans la plupart des programmes d'évaluation.

Les experts, des évaluateurs usagers ou des usagers évaluateurs ?

Dans les programmes d'évaluation, 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 : dans le domaine et thématiques du corpus d'une part, et dans les pratiques et usages en rapport avec les applications du système d'autre part.

Chaque juge est invité à examiner un ensemble de résultats au maximum, afin de lui 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 ou des limites des systèmes, mais constitue toutefois un référentiel qui ne garantit malheureusement pas la reproductibilité de l'expérience et par là l'obtention de résultats objectifs (Daille, 2002), il ne permet pas non plus d'évaluer le taux de silence dans certaines applications.

Enfin, des questions subsistent. Comment prendre en compte l'écart des *jugements cognitifs* voire *émotionnels* entre deux postures distinctes, celle d'un évaluateur dans le rôle d'usager potentiel et celle d'un usager chargé d'assumer la fonction d'évaluateur ? Comment s'assurer de la stabilité des jugements successifs d'un même sujet ? Dans un

protocole d'évaluation basé sur une expertise humaine, il est primordial d'analyser l'impact de subjectivité des experts et de corréler leurs appréciations pour ne pas biaiser le processus. Pour cela, la plupart des actions et campagnes d'évaluation ont eu recours au calcul des fameuses valeurs de corrélation telles que le coefficient de kappa, l'indice de Pearson ou le test de Khi2.

Les métriques ou l'approche quantitative, une solution qui pose problème

Dans plusieurs campagnes d'évaluation, l'usage de métriques comme approche quantitative est très courant et se présente comme une démarche scientifique rigoureuse, avec des outils de jugement et d'appréciation cadrés et normalisés. Cependant, il est très rare dans les recherches sur l'évaluation de ne pas voir surgir systématiquement le débat épistémologique entre les défenseurs et les opposants de l'approche quantitative.

D'après les défenseurs de l'évaluation quantitative, celle-ci permet de répondre au besoin croissant d'applications, manifesté par un foisonnement de méthodes et d'outils et devant lesquels l'usager éprouve ses difficultés d'évaluation et de sélection de manière objective. De leur côté, des opposants à l'approche quantitative estiment généralement que l'évaluation des systèmes est par essence de nature qualitative et que le versant quantitatif est nécessairement limité au cadre d'une validation plutôt qu'une évaluation. Si les arguments des deux camps sont recevables et traduisent deux approches complémentaires d'une même problématique, il est cependant important de noter que, grâce aux campagnes d'évaluation, l'approche quantitative de l'évaluation a non seulement permis de faire avancer l'état de l'art de manière significative, mais a également favorisé l'expansion du champ d'application des méthodes à base linguistique (Adda et al., 1999).

Dans une des campagnes étudiées (CESART), une des solutions prudentes pour atténuer les limites d'une évaluation quantitative, a été de demander à chaque juge d'établir un classement argumenté de l'ensemble des systèmes expertisés selon certains critères. Cet ordonnancement a permis de vérifier des questions sous-jacentes :

- Existe-t-il un accord inter-juges ? une différence significative dans l'ordonnancement des systèmes implique une divergence dans les regards des juges. Cette différence est-elle due au statut de juge responsable de l'expertise, doit-il être l'utilisateur final ou le spécialiste du domaine ?
- S'il y a un accord entre la majorité des juges, l'appréciation globale donnée par les experts est-elle en concordance avec les mesures calculées de manière algorithmique ?

L'évaluation orientée système, orientée usage

S'il y a un dénominateur commun à relever des différents programmes d'évaluation, il sera la réflexion permanente d'innover dans des métriques alternatives à celles

communément utilisées depuis des décennies (Nakache et Metais, 2005). Dans la plupart des travaux, l'évaluation est orientée système (*Modèle Cranfield*) dans le sens où elle est dépendante d'une métrique qui calcule l'écart entre les productions des systèmes et un référentiel préétabli ou parfois post édité dans le cas des *pooling method*. Or, la pertinence étant une fonction rétroactive, instable et évolutive, mais aucunement binaire, il devient légitime de se demander comment des métriques peuvent s'y prêter pour répondre à la satisfaction informationnelle des usagers et non à des considérations calculatoires des évaluateurs. Des recherches, notamment en sciences de l'information, ont montré les limites méthodologiques et théoriques de ce modèle et ont privilégié un changement de paradigme et une évaluation orientée usager (Chaudiron, 2002). Enfin, n'est-il pas intéressant d'orienter la réflexion vers un troisième paradigme, celui de l'*évaluation participative* (ou l'*évaluation de masse* ou *crowdrating*) avec ses propres théories, méthodologies et outils (popularité, communauté...) (Reber, 2013) ?

Pour des raisons de coût et de simplification, la plupart des campagnes d'évaluation ont été orientées systèmes et menées en mode *in vitro*, les interactions entre l'usager potentiel et son environnement naturel ont été délaissées de l'étude. Toutefois, la campagne INFILE sur le filtrage de l'information dans une perspective de veille, a été très prudente sur cette question et a intégré dans son protocole un maximum de vérités-terrain sur les pratiques de veille afin d'être le plus proche des conditions naturelles (mode *in vivo*).

CONCLUSION

Finalement, le traitement automatique de l'information relève à la fois de la démarche technologique et de la démarche scientifique. La perception de l'évaluation dans ce domaine varie considérablement entre les développeurs, les industriels, les chercheurs et les usagers potentiels... L'évaluation des systèmes informatiques ne doit plus être perçue seulement comme outil qui contribue à développer le volet technologique mais aussi en tant qu'un indicateur de progrès de la recherche scientifique. L'observation des dispositifs d'évaluation, l'analyse des pratiques d'évaluation et l'étude de cadres méthodologiques d'évaluation ouvrent de nouvelles perspectives de recherche scientifique dans le champ des sciences de l'information et de la communication, qu'on peut nommer abusivement "*l'évaluation de l'évaluation*".

Au regard de la place primordiale des Technologies de l'Information et de la Communication dans les travaux de recherche en SIC, la question de leur évaluation devient alors, elle aussi, primordiale. Il ne serait pas alors absurde de conclure que le problème épistémologique central de la question étudiée SIC est comment approcher les interactions usager-système dans un paradigme d'évaluation et introduire des modèles non contestables qui font référence à l'usage et aux pratiques d'une part et en concordance à des référentiels et métriques d'autre part.

Ce bilan épistémologique des recherches sur l’évaluation des systèmes de traitement automatique de l'information, laisse comprendre qu'en sciences de l'information aussi,

les relations entre théorie et pratique en matière d'évaluation demeurent une aporie opposant et reliant, à la fois, le *système technicien*, privé de sens et la recherche permanente d'une théorie, génératrice de sens (Figari, 2013 ; Rodriguez-Pabón, 2005).

RÉFÉRENCES BIBLIOGRAPHIQUES

Adda, G., Mariani, J., Paroubek, P., Rajman, M., Lecomte, J. (1999). L'action GRACE d'évaluation de l'assignation des parties du discours pour le français. *Revue Langues*, Vol. 2, No. 2, 119-129.

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

Chaudiron, S., Ihadjadene, M. (2002). Quelle place pour l'usager dans l'évaluation des SRI ? In *: Recherches récentes en sciences de l'information : Convergences et dynamiques*, 2002, Toulouse.

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

Figari, G. (2013). L'évaluation entre « technicité » et « théorisation » ? *Mesure et évaluation en éducation,* vol. 36 (3).

Guérin-Schneider, L., Tsanga Tabi, M. (2017). L'Analyse du Cycle de Vie, nouvel outil d'évaluation environnementale à l'appui des politiques publiques locales : Quelle appropriation par les services d'assainissement ? In : *Gestion et management public*, volume 5 / 4(2), 61-83.

Jorro, A., Droyer, N. (dir.) (2019). *L'évaluation, levier pour l'enseignement et la formation*. De Boeck Supérieur, 2019.

Le Priol, F. (2000). *Extraction et capitalisation automatiques de connaissances à partir de documents textuels*. Paris, Thèse de doctorat, Université Paris-Sorbonne.

Michel, C., Rouissi, S. (2003). Génération de documents d'évaluation des connaissances pour l'e-learning. In *6ème Colloque International sur le Document Électronique (CIDE6)*, Caen, 2003.

Nakache, D., Metais, E. (2005). Évaluation : nouvelle approche avec juges. In *Congrès Informatique des organisations et systèmes d'information et de décision (INFORSID'05)*, Grenoble, 2005.

Organisation internationale de normalisation, (1999). *Technologies de l'information - Évaluation de produits logiciels - Partie 1 : Aperçu général*. ISO/IEC 14598-1:1999.

Paroubek, P., Chaudiron, S., Hirschman, L. (2007). Principles of Evaluation in Natural Language Processing. *Revue Traitement Automatique des Langues (TAL)*, vol. 48, n° 1.

Pincemin, B. (1999). Construire et utiliser un corpus : le point de vue d'une sémantique textuelle interprétative. In Condamines, A., Péry-Woodley M.-P., et Fabre C. (dir), *Atelier Corpus et TAL : pour une réflexion méthodologique (TALN 99)*, Corse.

Popescu-Belis, A. (2007). Le rôle des métriques d'évaluation dans le processus de recherche en TAL. *Revue Traitement Automatique des Langues (TAL)*, vol. 48, n° 1.

Reber, B. (2013). Évaluation participative des technologies. In Casillo, I. et al. D. (dir.), *Dictionnaire critique et interdisciplinaire de la participation*. Paris, GIS Démocratie et Participation, 2013.

Reider, Harry R. (2000). *Benchmarking strategies a tool for profit improvement*. New York : John Wiley, 2000.

Rodriguez-Pabón, O. (2005). *Cadre théorique pour l'évaluation des infrastructures d'information géo-spatiale*. Université Laval. Thèse de doctorat, 2005.

Salton, G., McGill, M. (1983). *Introduction to Modern Information Retrieval*. Coll. Computer Science S., McGraw-Hill, 1983.

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

Timimi, I. (2006). L'évaluation des systèmes d'acquisition d'outils de terminologie : nouvelles métriques, nouvelles pratiques. *Colloque Jadt'06*, Besançon, avril 2006.

Van Rijsbergen, C.-J. (1979). *Information Retrieval, 2nd edition*. Butterworth-Heinemann.