Web sémantique et traitement automatique des langues

Bruno Menon

Web sémantique et traitement automatique des langues

Bruno Menon

1. Deux paradigmes disjoints?

Le Web sémantique (WS) et le traitement automatique des langues (TAL) sont deux champs qui sont susceptibles de modifier en profondeur le paysage des technologies de l'information et de la communication (TIC) dans un avenir proche. Bien qu'on ait l'intuition qu'il ne s'agit pas là de deux paradigmes complètement étrangers l'un à l'autre, il est plus facile d'identifier entre eux des points de friction que des points de contact.

Une enquête sur Google

Une rapide enquête sur Google (http://www.google.fr/) fait apparaître plus de 700 000 occurrences de la chaîne "Semantic Web", contre moins de 500 000 occurrences pour la chaîne "Natural Language Processing". La conjonction des deux termes de recherche donne, quant à elle, environ 10 000 résultats. Le Web sémantique est donc, malgré son apparition récente, davantage présent sur Internet que le TAL, discipline pourtant établie depuis plusieurs décennies. De plus, les pages mentionnant à la fois l'un et l'autre sont d'une relative rareté (2% ou moins). D'autres croisements de chacun de ces termes, avec par exemple "Artificial Intelligence" ou "Information Retrieval" sont systématiquement plus fructueux, comme le montre le graphique ci-dessous.

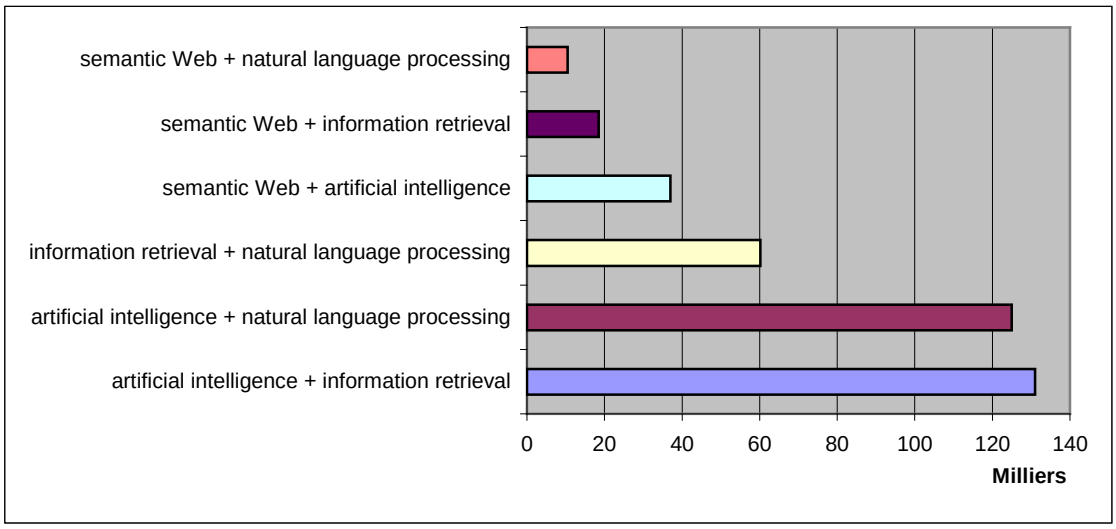

Fig. 1 **Une enquête sur Google**
Nombre de pages contenant les différentes paires de termes

Ce petit sondage, sans aucune prétention scientifique, semble malgré tout indiquer une faible interaction des deux domaines. Peut-être est-ce en partie une question de personnes, car les promoteurs du WS au World Wide Web Consortium (W3C) sont pour l'essentiel issus de l'univers des bibliothèques numériques ou de l'intelligence artificielle, mais pas du TAL.
Plus profondément, on peut lire ces projets comme davantages parallèles et concurrents que complémentaires. Les deux visions sont fondées sur des constats somme toutes similaires à propos des limitations actuelles des TIC, mais privilégient des voies sensiblement différentes pour les dépasser. C'est du moins l'hypothèse que l'on tentera d'illustrer ici.

Une explication de textes

A l'appui de cette hypothèse, on examinera deux textes qui, chacun à leur manière, définissent respectivement le projet du WS et celui du TAL.

[A] L'article très souvent cité de Scientific American, en 2001 : Tim Berners-Lee, James Hendler, Ora Lassila. *The Semantic Web*, Scientific American, May 2001. [Une traduction de cet article par Elisabeth Lacombe et Jo Link-Pezet est disponible à l'URL : http://www.urfist.cict.fr/lettres/lettre28/lettre28-22.html]

[B] Un texte du professeur Hans Uszkoreit, à l'université de la Sarre, proposant un panorama synthétique du TAL : Hans Uszkoreit. *What is computational linguistics?* [En ligne] http://www.coli.uni-sb.de/~hansu/what_is_cl.html (consulté le 13 mai 2004)

Tout d'abord, voici la phrase qui est couramment utilisée pour donner une brève définition de ce qu'est le Web sémantique :

- "The Semantic Web is an extension of the current web in which information is given well-defined meaning, better enabling computers and people to work in cooperation." [A]
  "Le Web sémantique est un extension du web actuel, dans laquelle l'information reçoit une signification bien définie, améliorant les possibilités de travail collaboratif entre les ordinateurs et les gens."

De cette première citation, on comprend que l'information présente sur le Web n'a pas de signification bien définie, à quoi le WS doit remédier. On lit plus loin :

- "The Semantic Web will enable machines to COMPREHEND semantic documents and data, not human speech and writings." [A]
  "Le Web sémantique rendra les machines à même de COMPRENDRE des documents et des données sémantiques, pas la parole et les écrits humains"

Quant aux ambitions du TAL :

- "Our long term goal is the deep understanding of human language." [B]
- "Notre objectif à long terme est la compréhension en profondeur du langage humain."

Ces extraits situent clairement le projet du WS dans une perspective opposée à celle du TAL, malgré une certaine ressemblance de surface. Pour le WS, la parole et les écrits humains sont dépourvus de sémantique, tout au moins de la sémantique nécessaire à la compréhension / appréhension par les machines.

On le voit, les vocables "comprendre", "sémantique" et "signification" ne font pas référence, dans le contexte de ces travaux, aux mêmes notions : sémantique linguistique d'un côté, une sémantique formelle et significations *calculables* de l'autre. En fait, le projet du WS tout entier est fondé sur le contraste entre une information interprétable par les humains, qui fait la substance du Web actuel, et une information interprétable par les machines, qui sera construite en parallèle et constituera le Web sémantique. La communication entre ces deux univers n'est pas évoquée. A l'inverse, la communication homme / machine occupe une place centrale dans la problématique du TAL :

- " We teach computers to communicate with people. […] Today's computers do not understand our language but computer languages are difficult to learn and do not correspond to the structure of human thought." [B]
  "Nous apprenons aux ordinateurs à communiquer avec les gens. […] Aujourd'hui, les ordinateurs ne comprennent pas notre langage, mais les langages informatiques sont d'apprentissage difficile et ne correspondent pas à notre structure de pensée."

On est donc en présence de deux raisonnements qui partent de propositions tout à fait compatibles, mais les hiérarchisent différemment, et qu'on peut expliciter ainsi :

[A] Pour les concepteurs du Web sémantique :

- Le Web actuel est conçu pour être interprété par des humains.
- Le langage humain, et en particulier l'information présente sur le Web, ne présente pas de signification pour les machines, car il est non-formel.
- Il est donc nécessaire d'ajouter au Web de l'information formelle,qui soit interprétable par les machines.

[B] Pour le TAL :

- Le langage humain ne présente pas de signification pour les machines.
- Les êtres humains ont en général des difficultés à maîtriser parfaitement les langages formels, seuls interprétables par les machines .
- Il est donc nécessaire de concevoir des systèmes informatiques capables de communiquer avec les êtres humains de façon non-formelle, en langage naturel.

Au-delà de cette divergence fondamentale sur les questions de sémantique et d'interprétabilité, les deux paradigmes proposent des réponses technologiques concurrentes à des préoccupations fonctionnelles et opérationnelles communes :

- "The vision of the semantic Web is to provide computer interpretable markup of the Web's content and capability, thus enabling automation of many tasks currently performed by human beings." [A]
  "La vision du Web sémantique est de fournir un marquage des contenus et des capacités du web qui soit interprétable par des machines, rendant ainsi possible l'automatisation de nombreuses tâches aujourd'hui accomplies par des êtres humains."

Les questions d'automatisation des traitements sont donc prépondérantes, comme celles de productivité des travailleurs du savoir :

- " Even if the language the machine understands and its domain of discourse are very restricted, the use of human language can increase the acceptance of software and the productivity of its users." [B]
  " Même si le langage compris par la machine et son univers de discours sont très restreints, l'usage du langage naturel peut accroître l'acceptation des logiciels et la productivité de leurs utilisateurs."

Enfin, c'est la manipulation de l'information non structurée qui constitue l'enjeu majeur pour les deux approches et appelle des développements appropriés :

- "The Semantic Web will bring structure to the meaningful content of Web pages." [A]
  "Le Web Sémantique appportera de la structure aux contenus significatifs des pages web."
- "The real power of the Semantic Web will be realized when people create many programs that collect Web content from diverse sources, process the information and exchange the results with other programs" [A]
  "La vraie puissance du Web Sémantique sera atteinte lorsque l'on aura créé de nombreux programmes pour recueillir l'information de diverses sources, traiter cette information et échanger leurs résultats avec d'autres programmes."
- "The whole world of multimedia information can only be structured, indexed and navigated through language. For browsing, navigating, filtering and processing the information on the web, we need software that can get at the contents of documents. Language technology for content management is a necessary precondition for turning the wealth of digital information into collective knowledge." [B]
  "L'univers entier de l'information multimédia ne peut être structuré, indexé et parcouru qu'à travers le langage. Pour parcourir, filtrer et traiter l'information sur le web, il faut des logiciels qui puissent atteindre le contenu des documents. Les technologies linguistiques de traitement des contenus sont une condition nécessaire à la transformation de l'information numérique en savoirs collectifs."

En résumé, on a là encore des raisonnements sous-jacents qui s'appuient sur des postulats similaires, mais concluent à des stratégies opposées :

[A] Pour les concepteurs du Web sémantique :

- La très grande majorité de l'information électronique est disponible sous forme non structurée.
- Les systèmes informatiques ont en général des difficultés à exploiter correctement l'information non structurée.
- Il est donc nécessaire d'ajouter de la structure à cette information pour que les systèmes informatiques à venir puissent l'exploiter correctement.

[B] Pour le TAL :

- La très grande majorité de l'information électronique est disponible sous forme non structurée.
- Les systèmes informatiques ont en général des difficultés à exploiter correctement l'information non structurée.
- Il est donc nécessaire de concevoir des systèmes informatiques capables d'exploiter correctement l'information non structurée.

Soulignons pour finir que ces deux paradigmes ont en commun une vision ambitieuse et exigeante de l'univers informationnel, vision à long terme et par certains côtés utopique, mais très mobilisatrice et porteuse de réalisations immédiates et opérationnelles plus restreintes.

2. Des apports mutuels

Ce sont ces réalisations qui font que Web sémantique et traitement automatique des langues peuvent s'apporter beaucoup sur le plan des technologies, des méthodes et des ressources, malgré leurs divergences majeures sur le plan des principes fondateurs et du projet d'ensemble.

TAL → WS : ce que le traitement automatique des langues peut apporter au Web sémantique

Le TAL est souvent défini par l'énumération des différentes tâches que les systèmes sont amenés à accomplir, que l'on peut grossièrement répartir en deux familles selon qu'elles sont destinées à la communication homme / machine ou à l'accroissement de la productivité des travailleurs du savoir :

- Fonctions d'interface
  - Reconnaissance de la parole
  - Synthèse de la parole
  - Traduction automatique
  - Repérage de l'information
  - Traitement de requêtes en langage naturel à des bases de données
  - Recherche de réponses à des questions.
- Fonctions de productivité
  - Traduction assistée
  - Classification automatique (supervisée ou non supervisée : catégorisation ou clustering)
  - Extraction d'informations
  - Indexation contrôlée, assistée ou automatique
  - Résumé automatique
  - Génération automatique de textes.

Toutes ces fonctions sont aujourd'hui couvertes par des systèmes industriels ou pré-industriels, à des niveaux divers de qualité et dans des domaines plus ou moins restreints. La plupart sont susceptibles de contribuer utilement à la mise en place du Web sémantique, que ce soit en fournissant les interfaces indispensables ou en proposant des outils de productivité pour le recueil des métadonnées ou la construction d'ontologies.

Des fonctions d'interface

On l'a vu, le Web sémantique se préoccupe assez peu des interfaces homme / machine, ce qui est compréhensible : les données formalisées du WS sont avant tout destinées à être manipulées par des logiciels et ne nécessitent que des systèmes de gestion pour leurs auteurs (éditeurs d'ontologies et de métadonnées) ; quant aux interfaces des applications de haut niveau (agents ou services Web intégrés), elles semblent actuellement hors du champ d'études, puisqu'elles ne différeront pas sensiblement de ce que nous connaissons aujourd'hui.

Ces positions appellent cependant deux remarques :

- Si les métadonnées du WS sont avant tout la matière première des traitements et des raisonnements opérés par les applications, on ne doit pas perdre de vue qu'une partie d'entre elles au moins est susceptible de servir au repérage de l'information. Les moteurs de recherche sémantiques font du reste partie des applications visées par le WS. Or les langages d'interrogation de méta-données au format RDF sont à l'heure actuelle très rudimentaires sur le plan de l'ergonomie, même s'ils sont suffisamment expressifs : les plus répandus, inspirés de SQL, ne sont guère accessibles qu'aux informaticiens [14]. Des interfaces de recherche sur les métadonnées sont donc souhaitables, et on aura le choix entre des interfaces de type formulaire et des interfaces en langage naturel. Ces dernières devraient s'imposer

dès lors que les modèles de métadonnées utilisés atteignent un certain degré de complexité. Bien que ce thème ne semble pas mobiliser fortement les chercheurs, ni les industriels, on peut citer le projet MOSES (http://www.hum.ku.dk/moses/), qui s'intéresse à une interrogation en langage naturel de RDF. et doit se terminer début 2005. Les nombreux travaux menés dans années quatre-vingt autour de la génération de requêtes SQL à partir de questions en langage naturel mériteraient sans doute d'être revisités dans ce nouveau contexte.

- N'oublions pas non plus que les dispositifs clavier / écran ne sont pas les seuls imaginables pour accéder au Web sémantique. Des interfaces vocales ou multimodales sont hautement probables et les technologies de traitement de l'oral (reconnaissance et synthèse de la parole) seront alors forcément sollicitées, en amont ou en aval d'autres techniques de TAL. [9]

Des fonctions de productivité

Parmi les facteurs qui vont conditionner le succès du Web sémantique, trois des plus critiques [3] ont vocation à être envisagés, au moins en partie, sous l'angle du traitement automatique des langues :

- La disponibilité de contenus dûment annotés.
- La disponibilité, le développement et l'évolution cohérente des ontologies.
- La prise en compte du multilinguisme.

Création de méta-données et construction d'ontologies

Le volume des ressources effectivemement pourvues de métadonnées est aujourd'hui dérisoire au regard des efforts de recherche et de développement consacrés à RDF et à ses outils d'accompagnement. Même si l'on nuance cette observation en en rappelant que l'avènement du WS passe par des applications à forte valeur ajoutée, dans des domaines spécifiques et pas (encore) par la "sémantisation" de l'ensemble des contenus de la Toile, la création de métadonnées et d'ontologies restent des opérations problématiques sur les plans économique, cognitif et organisationnel.

Des outils existent d'ores et déjà pour aider les auteurs à encoder des ontologies dans le respect des standards, à présent suffisamment stables (voir par exemple Protégé, http://protege.stanford.edu/), ou pour créer des métadonnées dans les modèles les plus courants (comme Dublin Core).

Mais l'assistance au recueil des métadonnées [1] ou à la préparation d'une ontologie [2] [17] impliquent le recours à une famille de technologies d'ingénierie linguistique regroupées sous l'appellation "extraction d'information". Elles pourraient trouver là des débouchés considérables et devenir l'application vedette (ou *killer app*) du couplage TAL / WS. [19]

Ces techniques, dont l'essor est largement dû à la série de conférences MUC (Message Understanding Conference), entre 1987 et 1998, recouvrent des tâches plus ou moins complexes :

- Identification de termes : recueillir la terminologie d'un corpus scientifique ou technique en identifiant des groupes de mots présentant certains patrons syntaxiques et en étudiant leur distribution dans le corpus.
- Repérage d'entités nommées : reconnaître des entités telles que des noms d'entreprise, des noms de personnes, des noms de lieux, des dates, etc. Les méthodes utilisées peuvent être basées sur un apprentissage statistique d'exemples, sur la présence d'indices comme les formules de civilité, les prénoms, etc., ou sur la recherche de patrons syntaxiques.
- Remplissage d'un formulaire (template) : trouver des caractéristiques d'un objet ; par exemple, pour un produit, trouver son nom, la société qui le fabrique, son prix, etc.
- Découverte de relations sémantiques : suggérer des relations entre concepts, par l'étude des propriétés distributionnelles des termes (apparitions dans des contextes similaires) ou à l'aide de marqueurs phraséologiques.

- Découverte de relations entre objets du monde : par exemple identifier, à partir de textes, les relations gènes / protéines ou symptômes / maladie.
- Description d'un événement ou scénario : donner les caractéristiques d'un événement dans un texte (objets impliqués et modalités de réalisation) ; un exemple souvent cité est celui des fusions / acquisitions d'entreprises.
- Résolution de co-références : repérer quand, dans un texte, il est fait référence plusieurs fois à une même entité, même si cette entité est nommée de façons différentes (pronoms personnels, acronymes, périphrases).

Identification de termes et repérage d'entités nommées sont particulièrement utiles dans les phases initiales de constitution d'une ontologie. Les tâches de découverte de relations peuvent aider à structurer les concepts de l'ontologie, et les fonctions de remplissage de formulaires à documenter des attributs sur les instances de concepts.
L'ensemble de ces techniques, diversement combinées, fournit des procédures pour la création semi-automatique des métadonnées. L'outillage qui paraît actuellement le plus avancé dans ce domaine est celui que propose Ontotext Lab, avec sa plateforme KIM (http://www.ontotext.com/) [20]
Enfin, une technologie parente, peu développée car victime du succès de l'indexation en texte intégral, l'indexation automatique contrôlée (indexation de textes par des assignation de descripteurs issus d'une liste ou d'un thésaurus) pourrait faciliter la production de métadonnées thématiques fondées sur une ontologie (voir par exemple Machine Aided Indexing : http://mai.larc.nasa.gov/what.html).

Multilinguisme

Le Web est éminemment multilingue, ce que ne doit pas masquer la prépondérance de l'anglais dans les ressources en ligne. Le Web sémantique postule des représentations abstraites formalisées, au niveau conceptuel, de ces ressources. Ces représentations sont donc censées être indépendantes de la langue, linguistiquement neutres. Il n'en reste pas moins que certains constituants des métadonnées et des ontologies ressemblent furieusement à des éléments des langues naturelles, à des termes.
De plus, quel que soit le niveau d'automatisation des tâches, les différents systèmes produiront *in fine* des résultats qu'il faudra présenter, et, dans le cas du WS, justifier, aux utilisateurs. Il n'est pas déraisonnable d'imaginer que les usagers du WS voudront être à même d'interagir dans leur langue maternelle avec ces systèmes, et que les créateurs de métadonnées et d'ontologies auront un travail plus productif et de meilleure qualité s'ils n'ont pas à affronter des barrières linguistiques [3].
Tout cela suppose la disponibilté de ressources linguistiques multilingues, de procédures de fusion et de présentation de métadonnées libellées dans des langues différentes (voir par exemple, MUMIS, un projet de traitement de l'information multilingue et multimédia : http://parlevink.cs.utwente.nl/projects/mumis/), et, à des degrés divers, de systèmes d'aide à la traduction.

Une intégration WS / TAL

Après ce tour d'horizon, on peut proposer un schéma du Web sémantique où les fonctions assurées par des composants de traitement automatique des langues seraient intégrées.
quelque peu amendé par rapport aux présentations habituelles qui sont faites de la "chaîne alimentaire" du WS : la présence de ces fonctions y est au mieux implicite, et au pire jugée superflue.

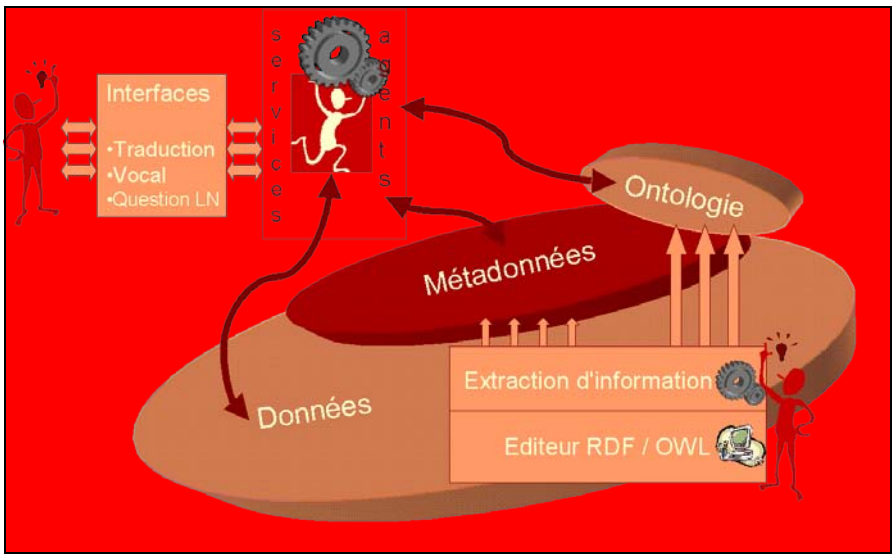

Fig. 2 **Web sémantique avec TAL intégré**

WS → TAL : ce que le Web sémantique peut apporter au traitement automatique des langues

Des standards

L'interopérabilité des différentes ressources disponibles en ligne joue un grand rôle dans la conception du Web sémantique. La condition nécessaire – bien que non suffisante – à cette intéropérabilité est la définition de standards permettant d'offrir des langages formels communs à toutes les applications du WS.

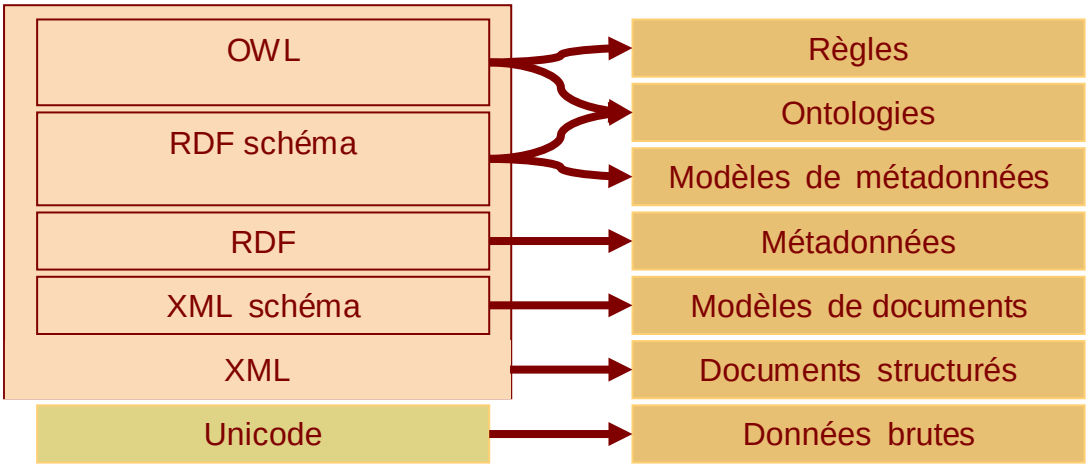

Fig. 3 **Les standards du Web sémantique**

Mis en avant par le W3C, ces standards commencent à se diffuser assez largement. Dans la communauté du traitement automatique des langues, les besoins d'interopérabilité et de réutilisation se font sentir de longue date, mais le domaine reste aujourd'hui peu normalisé. Si XML s'y répand rapidement comme format standardisé d'encodage et d'échange des données, RDF ou OWL sont moins fréquemment utilisés. Un intérêt certain pour ces formalismes s'amorce toutefois. On note par exemple la présence d'un atelier "Natural Language Processing and XML", dédié à ce thème, à la conférence de l'ACL (Association for Computational Linguistics, voir en ligne http://www.ling.helsinki.fi/~gwilcock/NLPXML-2004/).

Le TAL produit en effet des ressources spécifiques, en particulier des corpus annotés et des lexiques, pour lesquelles il est important de disposer de métadonnées fiables et normalisées. Notons qu'en ce qui concerne les corpus, RDF peut être utilisé pour enregistrer les descriptions de ces corpus, mais surtout pour encoder les annotations linguistiques qui leur sont ajoutées. Des projets tels que "ISLE Meta Data Initiative" travaillent dans ce sens [5]. Le sous-comité de l'ISO (International Organization for Standardization, http://www.tc37sc4.org) chargé des ressources linguistiques est également favorable à l'utilisation de ces standards.
Les dictionnaires électroniques sont des lexiques sur lesquels s'appuient de nombreux systèmes de TAL, pour lesquels ils constituent une ressource cruciale. Leur complexité et leur coût font de leur normalisation un enjeu critique pour le secteur. La normalisation des formalismes de dictionnaires électroniques a fait l'objet de nombreux travaux dans les années 90, avec de multiples propositions qui n'ont pas recueilli le consensus nécessaire à leur généralisation (GENELEX, MULTILEX, ACQUILEX, etc.) Certains de ces travaux se prolongent aujourd'hui vers la définition de schémas RDF pour l'encodage des dictionnaires électroniques [7].

Des métadonnées

De nombreuses applications de TAL, reposent sur des méthodes d'apprentissage. Mais ces méthodes nécessitent la constitution préalable de corpus étiquetés, ce qui est souvent un obstacle. Dans certains cas, un sous-ensemble des métadonnées produites pour des applications de WS pourrait être utilisé à ces fins, ce qui épargnerait une partie du travail d'étiquetage. Il semble en outre que cet apprentissage soit plus efficace, notamment en classification automatique, lorsque les documents du corpus d'entraînement sont accompagnés de métadonnées détaillées [8].
De plus, les techniques de classification peuvent être appliquées avec profit sur des collections de ressources pourvues d'annotations [22], *a fortiori* de métadonnées formalisées correctement associées à leur schéma et à leur ontologie [16], en particulier lorsque les entités à classer sont hétérogènes : collections mêlant texte et images, classement d'artefacts. Enfin, l'utilisation des métadonnées pour suppléer ou compléter les mots du texte intégral lors de l'indexation des ressources peut améliorer le comportement des moteurs de recherche [10] [21].

Des ontologies

Les applications de TAL en vraie grandeur butent souvent sur la question des ressources sémantiques nécessaires à des performances satisfaisantes. Thésaurus (comme MeSH), réseaux lexicaux (comme WordNet et ses versions européennes), réseaux sémantiques (comme UMLS) ont souvent été mis à profit par divers systèmes, mais leur intégration ne va pas sans demander des investissements importants. Le fait de disposer d'ontologies adéquates au domaine traité, encodées dans un format normalisé et dont le modèle est formellement explicité sera un facteur déterminant de la qualité des résultats obtenus et de la viabilité économique des applications de TAL.
Adossées aux mêmes ontologies que les applications de WS auxquelles elles seront intégrées, elles n'en rempliront que mieux les fonctions d'interface et de productivité évoquées plus haut. On voit ainsi s'amorcer une sorte de cercle vertueux où des systémes de TAL contribuent à l'alimentation d'ontologies pour le WS, mais les exploitent aussi pour affiner leurs analyses, améliorant ainsi les résultats obtenus à l'itération suivante.

TAL ↔ WS : des synergies indéniables

La maturité technologique du Web sémantique est pratiquement acquise, et le socle des standards y contribue incontestablement. Sa maturation opérationnelle et économique va dépendre de la capacité des entreprises à constituer une masse critique de données formalisées (ontologies, métadonnées). Mais les coûts initiaux de tels développements sont une barrière pour beaucoup d'entreprises. L'apport de l'ingénierie linguistique peut contribuer

à la réduction de ces coûts, et favoriser ainsi le décollage d'applications de WS à l'échelon industriel.
De son côté, l'ingénierie linguistique, malgré une histoire déjà longue et d'incontestables succès, n'a pas connu l'expansion fulgurante qu'on lui prédisait. Ces difficultés sont sans doute dues en partie au manque d'interopérabilité des différents outils disponibles, à une modularité insuffisante des systèmes, et au caractère faiblement cumulatif des développements techniques. L'infrastructure du WS apporte un début de réponse à ces problèmes, avec un cadre normatif bien documenté, largement diffusé et appuyé par des outils dont beaucoup sont des logiciels libres (ou *open source*).
A cela s'ajoutent des opportunités de mutualisation au moins partielle des coûts de représentation des connaissances, avec les possibilités d'ontologies partagées.
D'incontestables synergies sont donc présentes, et il appartiendra aux acteurs des deux domaines d'en tirer parti. Le partenariat entre les sociétés Mondeca, entreprise française très présente dans les travaux autour du Web sémantique (http://www.mondeca.com/), et Temis, acteur majeur dans le domaine de la fouille de texte (http://www.temis-group.com/), ou l'existence de sociétés comme Ontotext Lab, dont l'un des objectifs est l'intégration des deux technologies, sont davantage l'exception que la règle aujourd'hui, mais nous font entrevoir ce que pourrait être un WS^2, un Web doté à la fois de sémantique formelle et de sémantique linguistique.

Références

[1] *Annotation for the Semantic Web*. Ed. Handschuh S., Staab S. Amsterdam : IOS Press ; 2003. (Frontiers in Artificial Intelligence and Applications ; 96)

[2] Aussenac-Gilles N., Bourigault D. *Construction d'ontologies à partir de textes.* In TALN (10 ; 2003 ; Batz-sur-Mer, France). [En ligne] http://www.univ-tlse2.fr/erss/membres/bourigault/TALN03-tutoriel-bourigault-aussenac.doc (consulté le 13 mai 2004)

[3] Benjamins V. R., Contreras J., Corcho O., Gomez-Perez A. *Six Challenges for the Semantic Web.* In KR2002 : Semantic Web workshop (8 ; 2002 ; Toulouse, France). [En ligne] http://www.isoco.com/isococom/whitepapers/files/SemanticWeb-whitepaper-137.pdf (consulté le 19 mai 2004)

[4] Bontcheva K., Cunningham H. *The Semantic Web: A New Opportunity and Challenge for Human Language Technology.* In International Semantic Web Conference : Workshop on Human Language Technology for the Semantic Web and Web Services (2 ; 2003 ; Sanibel Island, Florida, USA). [En ligne] http://gate.ac.uk/sale/iswc03/iswc03.pdf (consulté le 12 mai 2004)

[5] Broeder D., Wittenburg P., Sloman, B. *EAGLES/ISLE: A Proposal for a Meta Description Standard for Language Resources*. White Paper. LREC Workshop (2000 ; Athens). [En ligne] http://www.mpi.nl/world/ISLE/documents/papers/white_paper_11.pdf (consulté le 19 mai 2004)

[6] Buitelaar P., Calzolari N., Declerck T., Lenci A. *Towards a Language Infrastructure for the Semantic Web.* In International Semantic Web Conference : Workshop on Human Language Technology for the Semantic Web and Web Services (2 ; 2003 ; Sanibel Island, Florida, USA). [En ligne] http://www.dfki.de/~paulb/iswc03-hltsw.pdf (consulté le 12 mai 2004)

[7] Calzolari N., Ide N., Lenci A.. *RDF Instantiation of ISLE/MILE Lexical Entries.* In ACL 2003 : Workshop on Linguistic Annotation: Getting the Model Right (41 ; 2003 ; Sapporo, Japon). [En ligne] http://acl.ldc.upenn.edu/W/W03/W03-1905.pdf (consulté le 12 mai 2004)

[8] Dini L. *NLP Technologies and Semantic Web: Risks, Opportunities and Challenges.* Intelligenza artificiale 2004 ; 1(1) : 67-71. [En ligne] http://www.celi.it/dini_aiia.PDF (consulté le 13 mai 2004)

[9] Dorai G. K., Yacoob Y. *Facilitating Semantic Web Search with Embedded Grammar Tags.* In IJCAI-01 : Workshop on E-Business & the Intelligent Web (16 ; 2001 ; Seattle, USA). [En ligne] http://www.csd.abdn.ac.uk/ebiweb/papers/dorai.pdf (consulté le 12 mai 2004)

[10] Finin T., Joshi A., Shah U. *Information retrieval on the semantic web.* In International Conference on Information and Knowledge Management (11 ; 2002 ; McLean, Virginia, USA). New York : ACM Press ; 2002. [En ligne] http://www.umbc.edu/~finin/papers/cikm02/cikm02.pdf (consulté le 13 mai 2004)

[11] Fürst, F. *L'ingénierie ontologique.* Rapport de recherche 02-07 : Laboratoire de recherche en informatique de Nantes ; 2002. [En ligne] http://www.sciences.univ-nantes.fr/info/perso/permanents/furst/papers/RR02-07.pdf (consulté le 12 mai 2004)

[12] Jokinen K., Wilcock G. *Generating Responses and Explanations from RDF/XML and DAML+OIL.* In IJCAI-2003 : Knowledge and Reasoning in Practical Dialogue Systems (18 ; 2003 ; Acapulco, Mexique). [En ligne] http://www.ling.helsinki.fi/~gwilcock/Pubs/IJCAI-03.pdf (consulté le 13 mai 2004)

[13] Katz, B., Lin, J., Quan, D. *Natural language annotations for the Semantic Web*. In International Conference on Ontologies, Databases, and Application of Semantics (1 ; 2002 ; Irvine, California, USA). Heidelberg : Springer-Verlag ; 2002. (Lecture Notes in Computer Science ; 2519). [En ligne] http://www.ai.mit.edu/projects/infolab/publications/Katz-etal-ODBASE02.pdf (consulté le 13 mai 2004)

[14] Kavalec M., Labsky M., Svab O., Svatek V. *Querying the RDF: Small Case Study in the Bicycle Sale Domain.* In Workshop on Databases, Texts, Specifications and Objects (2004 ; Ostrava, République tchèque). [En ligne] http://nb.vse.cz/~svatek/dtso04.pdf (consulté le 13 mai 2004)

[15] Klein E., Potter S. *An ontology for NLP services.* In LREC 2004 : Workshop on a Registry of Linguistic Data Categories within an Integrated Language Resource Repository Area (4 ; 2004 ; Lisbonne, Portugal).
[En ligne] http://www.ltg.ed.ac.uk/~ewan/Papers/FTP/Klein:2004:ONS.pdf (consulté le 13 mai 2004)

[16] Maedche A., Zacharias V. *Clustering Ontology-based Metadata in the Semantic Web.* In Principles of Data Mining and Knowledge Discovery (6 ; 2002 ; Helsinki, Finlande). Heidelberg : Springer-Verlag ; 2002. (Lecture Notes in Computer Science ; 2341) [En ligne] www.fzi.de/KCMS/kcms_file.php?action=link&id=36 (consulté le 19 mai 2004)

[17] Maedche A. *Ontology Learning for the Semantic Web.* Boston : Kluwer ; 2002 (The Kluwer International Series in Engineering and Computer Science ; 665)

[18] Motta M., Vargas-Vera M. *AQUA – Ontology-Based Question Answering System.* In MICAI 2004: Advances in Artificial Intelligence (3 ; 2004 ; Mexico, Mexique). Heidelberg : Springer-Verlag ; 2004. (Lecture Notes in Computer Science ; 2972)

[19] Poibeau T. *Extraction automatique d'information : Du texte brut au web sémantique.* Paris : Hermès Science ; 2003.

[20] Popov B. et al. Towards Semantic Web Information Extraction. In International Semantic Web Conference : Workshop on Human Language Technology for the Semantic Web and Web Services (2 ; 2003 ; Sanibel Island, Florida, USA). [En ligne] http://www.ontotext.com/publications/SemAIR_ISWC169.pdf (consulté le 19 mai 2004)

[21] Schmitz-Esser W. *Meaning, understanding and the organization of knowledge in a multilingual world - New tools for new tasks : Ontologies.* In Linguistic Cultural Identity and International Communication, Ed Vielberth J. Drexel G. Saarbrück : AQ-Verlag ; 2003.

[22] Vignolet L., Denoue L. *L'importance des annotations : application à la classification des documents du Web.* Documents Numériques 2000 ; 4 (1-2) : 37-57) [En ligne] http://www.fxpal.com/people/denoue/publications/docnum.pdf (consulté le 19 mai 2004)