Revue québécoise de linguistique

Présentation : TALN, Web et corpus

Louisette Emirkanian and Christophe Fouqueré

TALN, Web et corpus
Volume 32, Number 1, 2003

URI: https://id.erudit.org/iderudit/012241ar
DOI: https://doi.org/10.7202/012241ar

See table of contents

Publisher(s)

Université du Québec à Montréal

ISSN

0710-0167 (print)
1705-4591 (digital)

Explore this journal

Cite this document

Emirkanian, L. & Fouqueré, C. (2003). Présentation : TALN, Web et corpus. *Revue québécoise de linguistique*, *32* (1), 7–9. https://doi.org/10.7202/012241ar

Revue québécoise de linguistique, vol. 32, nº 1, 2003, © RQL (UQAM), Montréal

PRÉSENTATION : TALN, WEB ET CORPUS

Le Web peut être considéré à la fois comme un champ d'étude, une ressource et une base de données. Ces aspects liés au Web, faisant appel à des domaines divers de la linguistique, ont fait l'objet d'un colloque[1] les 26 et 27 novembre 2002 à Saint-Denis, France, dont une sélection de neuf articles paraît dans ce numéro.

Il est possible d'assimiler le Web à un simple corpus linguistique. Cette conception du Web est fondée sur deux de ses principales caractéristiques. D'une part, il apparaît comme un fond documentaire réellement représentatif des langues contemporaines au regard de sa masse considérable (en expansion perpétuelle), de l'extrême variété de ses textes (tant sur le plan thématique que sur le plan stylistique) et de la diversité sociologique de ses utilisateurs. D'autre part, il s'agit d'un médium spécifique sur plus d'un point, du fait de ses multiples fonctionnalités. Il en résulte de nouvelles pratiques qui ont d'importantes répercussions tant dans le domaine de la linguistique de corpus que dans celui du repérage ou de l'extraction d'informations. Mais cette utilisation du Web comme corpus soulève un certain nombre d'interrogations : le Web est-il vraiment une bonne source de données textuelles? Ces données peuvent-elles servir de base à une étude linguistique? Comment récupérer des données, quels sont les outils disponibles, comment traiter ces données, ces données sont-elles directement exploitables?

C'est dans cette perspective que se situent nombre des travaux actuels en traitement automatique du langage. Ainsi, dans l'article de **Eggert, Maurel et Piton**, la validation des règles de formation des gentilés a été effectuée après recherche directe d'exemples et de contre-exemples sur le Web. De même, **Fourour et Morin** montrent l'intérêt du Web comme source de données dans

1 Ce colloque a été organisé dans le cadre d'un projet de coopération Québec-France, financé par le ministère français des Affaires Étrangères, le ministère québécois des Relations Internationales, et a reçu le soutien de l'université de Paris XIII, de la Maison des Sciences de l'homme de Paris-Nord, du Laboratoire de linguistique informatique et du Laboratoire informatique de Paris-Nord. Nous tenons à remercier plus particulièrement messieurs Steeve Harbour, Pierre Moeglin et François Payeur pour leur soutien constant.

une application de reconnaissance semi-automatique des entités nommées. Le Web sert alors de vaste corpus encyclopédique : si le contexte local d'une entité nommée ne permet pas sa classification, le Web peut permettre de trouver d'autres exemples d'emploi permettant cette catégorisation. **Hathout et Tanguy**, quant à eux, s'intéressent plus spécifiquement à la détection et à l'analyse d'unités lexicales construites par suffixation ou préfixation. Là encore, le Web est largement mis à contribution comme base d'exemples.

La démarche n'est toutefois pas exempte de ses propres insuffisances. Ainsi, il n'y a pas un unique niveau de langue sur le Web : se côtoient des articles de journaux, des présentations de personnes ou de produits, des messages plus ou moins transcrits de l'oral. Qui plus est, les documents sont peu ou mal formatés. Leurs prétraitements sont dès lors inévitables. Ceux-ci sont abordés sous divers aspects dans plusieurs articles. Hathout et Tanguy explicitent les différentes étapes nécessaires pour que les occurrences d'unités lexicales soient exemptes d'erreurs (typographiques principalement). **Namer** effectue, elle aussi, une série de prétraitements en vue de constituer, à partir du Web, une base de données morphologiques. Bien au delà d'une simple nomenclature, l'automatisation tout autant que la taille du corpus permettent de constituer des données répondant à des critères fins : précision sur l'environnement lexical ou syntaxique des éléments recherchés, choix d'un procédé de formation particulier, etc.

Avec des objectifs très différents, sur lesquels nous reviendrons, l'article de **Fouqueré et Issac** et celui de **Emirkanian et Chieze** exposent les démarches permettant effectivement de constituer un corpus à partir du Web et donnent des indications sur la pertinence informationnelle et thématique des documents obtenus. L'utilisation du Web comme source de données passe en effet par une analyse de la fiabilité de ces documents, et par une normalisation du contenu. Les opérations qui doivent être mises en place sont nécessaires mais délicates : toute modification de la source entraîne un biais dans l'analyse ou l'utilisation des données. La quantification de ces facteurs apporte quelques éléments de réponse, toutefois la diversité du Web est telle que ceux-ci peuvent sans doute difficilement être généralisés.

Le Web peut aussi être utilisé comme vecteur d'informations linguistiques. C'est ce que montre **Pierrel** dans un article consacré aux ressources lexicales du laboratoire Analyse et traitement informatique de la langue française (ATILF) accessibles via le Web. De la base de documents que constitue Frantext aux outils d'exploitation (de type morphologique entre autres), le Web est alors un médium entre l'utilisateur et le corpus.

Sa facilité d'accès, sa disponibilité en font un outil précieux qu'il serait regrettable qu'un linguiste ignore. La normalisation des documents présentés évite le recours aux outils de pré- et posttraitement mentionnés ci-dessus.

Par ailleurs, c'est la problématique du questionnement du Web qui est l'élément central des autres articles qui composent ce numéro. Emirkanian et Chieze, Fouqueré et Issac ont étudié les mécanismes qui devraient être mis en œuvre dans le cas de reformulation de requêtes. En effet, le nombre de documents disponibles sur le Web est tel que trouver la bonne information tient de la gageure pour chacun d'entre nous, y compris avec les meilleurs moteurs de recherche actuels. Leur étude, restreinte aux cas de structures locatives, montre divers aspects de cette problématique : variabilité des documents extraits du Web, importance de la prise en compte, même partielle, de la structure argumentale, de l'enrichissement morphologique et synonymique, enfin du contexte. Dans un autre registre, **Buvet, Moreau et Silberztein** montrent comment une désambiguïsation des substantifs polysémiques d'un document est possible, en se servant de grammaires locales et de postanalyse sur des documents-réponses. Enfin, c'est par le biais d'une typologie des réponses et des requêtes que **Benamara et Saint-Dizier** développent un logiciel facilitant la tâche de l'utilisateur dans ses démarches dans un système coopératif en langue naturelle.

Il convient de préciser que toutes les recherches dont il est fait état ici portent sur le Web en langue française. Cette partie ne représente que 1% du Web mondial, il est toutefois primordial que de telles études soient menées.

Louisette Emirkanian
Université du Québec à Montréal
Christophe Fouqueré
Université de Paris XIII