De l'exemple construit à l'exemple attesté : un système de requêtes syntaxiques pour non-spécialistes

Ilaine Wang[1] Sylvain Kahane[1] Isabelle Tellier[2]
(1) MoDyCo (UMR 7114), CNRS, Université Paris Ouest Nanterre La Défense
(2) LaTTiCe (UMR 8094), CNRS, ENS Paris, Université Sorbonne Nouvelle - Paris 3,
PSL Research University, USPC (Université Sorbonne Paris Cité)
i.wang@u-paris10.fr, sylvain@kahane.fr, isabelle.tellier@univ-paris3.fr

RÉSUMÉ

Notre objectif est de permettre aux apprenants de langue d'accéder aux données présentes dans un corpus en facilitant la formulation de requêtes portant sur des critères syntaxiques. Nous proposons une méthodologie utilisant des mesures de similarité classiques pour comparer des séquences d'étiquettes obtenues par des programmes d'annotation.

ABSTRACT

From built examples to attested examples : a syntax-based query system for non-specialists

Our purpose is to allow non-specialists like language learners to access corpora, making syntactic queries easy to express. We propose a methodology including a syntactic parser and using common similarity measures to compare sequences of (automatically produced) morphosyntactic tags.

MOTS-CLÉS : linguistique de corpus, système de requête, construction syntaxique.

KEYWORDS: corpus linguistics, query system, syntactic construction.

1 Introduction

Depuis l'ère numérique, le développement de la linguistique de corpus s'accompagne du développement de systèmes de requêtes permettant d'exploiter les corpus comme des ressources linguistiques. De la même façon que la recherche d'information s'est dotée de moteurs de recherche, la linguistique de corpus s'est équipée d'outils tels que les concordanciers, qui reposent sur des requêtes à base de mots-clés. L'exploitation d'un corpus n'est pas seulement rendue possible par l'utilisation de concordanciers, elle est également déterminée par ces derniers : ce que l'on tire des corpus dépend fortement des possibilités offertes par les outils qui les exploitent (Anthony, 2013), et reposer sur des mots-clés peut être une contrainte pour qui s'intéresse à certaines constructions complexes et/ou qui n'ont pas de marqueur lexical précis. On considérera par exemple le cas des propositions relatives, marquées non pas par un item lexical mais par la catégorie grammaticale des pronoms relatifs.

Il n'est possible aujourd'hui de rechercher des structures complexes qu'en les décrivant précisément ce qui suppose de connaître le langage de requête et le jeu d'étiquettes du corpus annoté. Ces connaissances, communes en linguistique outillée et en TAL, requièrent des efforts importants de la part de non-spécialistes tels que les apprenants ou enseignants de langue. Dans la suite de l'article, nous commençons par faire le tour des besoins et des outils actuellement disponibles pour l'utilisation

des corpus en didactique des langues. Nous proposons ensuite une chaîne de traitements prenant en considération les difficultés éventuelles de non-spécialistes et comprenant un système de requête fondé sur une notion de similarité syntaxique.

2 Interrogation de corpus

Les apprenants et les enseignants de langues ne sont généralement pas des linguistes. Ils sont rarement initiés aux méthodes de la linguistique outillée ou du TAL alors qu'ils sont de plus en plus nombreux à percevoir l'intérêt d'accéder à des corpus. Après avoir évoqué les tenants et les aboutissants de l'accès à des données attestées en didactique des langues, nous présentons les outils disponibles pour interroger des corpus, en montrant leurs limites, notamment quand la requête porte sur une construction syntaxique.

2.1 Utilisation des corpus en didactique des langues

Les corpus de locuteurs natifs représentent en didactique des langues une ressource intéressante puisqu'ils constituent pour l'enseignant comme pour l'apprenant des ensembles de documents authentiques dans lesquels il est possible d'observer ce qui est considéré comme naturel ou usuel dans la langue cible (voir notamment les travaux de Chambers (2005, 2010) et de Cavalla (2015) pour l'aide à l'écrit en Français Langue Etrangère). Cette exposition à des données authentiques peut être indirecte (distribution en classe de concordances réalisées au préalable par exemple) ou bien être l'aboutissement d'une démarche plus directe. La seconde méthode est particulièrement exploitée dans l'approche que Johns nomme *Data-Driven Learning* (DDL), et qui considère l'apprenant comme un « chercheur dont l'apprentissage devrait être motivé par l'accès à des données linguistiques » [1] (Johns, 1991, p.2). L'apprenant doit être actif dans son apprentissage, être capable de formuler des hypothèses, d'observer et d'analyser des données langagières pour confirmer ou infirmer lui-même ses hypothèses, et enfin d'en formuler de nouvelles au besoin.

Or, dans la pratique, les apprenants peuvent considérer que les bénéfices évidents d'une confrontation directe à des corpus authentiques ne valent pas les efforts fournis ni le temps employé pour apprendre à utiliser de manière appropriée les outils d'exploration des corpus. Boulton (2012) cite en effet parmi les points négatifs soulevés par ses étudiants la complexité de l'interface de requête et donc la nécessité de recevoir une formation spécifique pour pouvoir exploiter au mieux les corpus. C'est à partir du même constat que Falaise *et al.* (2011) proposent un outil d'exploration de corpus arborés avec une interface plus simple, minimaliste (options cachées) et conviviale (interface graphique et non textuelle), qui n'empêche pas des requêtes fines et précises. Si cette simplification de l'interface permet effectivement de réduire de manière significative le temps nécessaire à la maîtrise de l'outil, elle présuppose néanmoins les mêmes connaissances que précédemment de la part de l'utilisateur.

2.2 Méthodes actuelles de requêtage en linguistique de corpus

Une des méthodes les plus courantes en linguistique de corpus est l'utilisation des concordanciers. Ces derniers, de plus de plus utilisés en didactique des langues, incluent tous au moins deux fonctionalités

1. Dans le texte original : *research workers whose learning needs to be driven by access to linguistic data*".

principales qui ont pour unité de base le mot : d'un côté, le calcul de statistiques mettant en évidence les propriétés du texte (ou corpus) étudié (nombre d'occurrences, distribution, collocation etc.), et de l'autre, les concordances KWIC (KeyWord In Context) où l'on retrouve le mot ou la séquence de mots cibles alignés et dans leur contexte originel. On peut remarquer qu'à la différence des requêtes dans les moteurs de recherche, les suites de mots données en entrée à un concordancier sont généralement des n-grams, soit des séquences de mots strictement contiguës, dont l'ordre est préservé. L'implémentation de *skipgrams* (n-grams non contigus) dans des logiciels de concordance est plus rare mais on note qu'il existe des outils de recherche d'unités phraséologiques à visée pédagogique, dont ConcGram et le Lexicoscope pour le français, qui l'autorisent. Ces derniers tiennent compte des variations de position et de dépendance à l'intérieur d'un syntagme grâce à un système prenant en entrée plusieurs mots [2] dits pivots, soit directement donnés par l'utilisateur, soit associés de manière itérative à partir d'un premier pivot (ou deux pour ConcGram) auquel on adjoint jusqu'à quatre mots co-occurrents repérées par l'outil (Cheng *et al.*, 2006; Kraif & Diwersy, 2012).

Il est toutefois possible de se libérer du mot et d'avoir directement recours aux étiquettes morpho-syntaxiques. En effet, l'appariement de deux segments comme "*la personne que je vois*" et "*ce rêve dont tu parles*" qui n'ont aucune unité lexicale commune mais qui, en revanche, partagent la même structure syntaxique, ne peut être réalisé qu'avec le patron "`DET NOM PROREL PRO VERBE`". Pour formuler une telle requête, l'utilisateur doit non seulement connaître ce jeu d'étiquettes mais également avoir suffisamment de connaissances linguistiques, notamment pour pouvoir associer à un mot la bonne partie du discours. Les expressions régulières permettent une expressivité encore plus grande, mais au prix d'une initiation encore plus poussée. L'outil GrETEL (Augustinus *et al.*, 2012) résoud en partie le problème puisqu'il offre la possibilité d'interroger un treebank en transformant automatiquement un exemple de structure syntaxique en requête, à l'instar de ce que nous proposons. Il permet ainsi à ses utilisateurs de s'affranchir de l'apprentissage d'un langage de requête complexe, mais s'adresse toutefois bien à des linguistes conscients de ce qu'ils recherchent et donc capables de paramétrer la requête en ce sens.

Nous nous attachons à aller plus loin dans l'ouverture des outils d'exploration de corpus en proposant une chaîne de traitements qui comprenne à la fois (1) la réduction de la complexité de l'interface du système de requêtes et (2) la réduction de la profondeur et de la variété des connaissances sollicitées de la part de l'utilisateur. Pour notre problématique, nous avons choisi pour le moment de n'exploiter que les parties du discours, sans prendre en compte la structure arborescente des *treebanks*.

3 Méthodologie

3.1 Chaîne de traitements

Notre objectif étant de simplifier au maximum la tâche de la formulation de la requête pour le non-spécialiste, nous proposons une méthodologie qui permettrait de passer "directement" d'un exemple donné en langage naturel à d'autres exemples illustrant la même construction syntaxique. Toutes les étapes de transformation et de comparaison des données seraient assumées par des traitements automatisés et ne solliciteraient donc pas plus de connaissances que celles nécessaires à la validation (ou l'invalidation) des résultats donnés en sortie. La chaîne de traitements complète détaillée en

2. Par mot, on entend ici le mot tel quel (forme fléchie) ou bien le lemme qui lui correspond, permettant aux utilisateurs de considérer ou non les variations morphologiques.

Figure 1 avec un exemple de proposition relative s'articule autour des étapes suivantes :

1. l'analyse (morpho)syntaxique automatique du (ou des) segment(s) donné(s) en entrée par l'utilisateur [3] ;
2. la transformation de l'input en langage naturel en une requête interprétable par la machine ;
3. la mesure de similarité syntaxique entre la requête et les phrases du corpus ;
4. la proposition à l'utilisateur des segments similaires regroupés en clusters ;
5. la sélection par l'utilisateur de l'exemple qui lui paraît le plus proche de sa requête, permettant ainsi d'affiner la requête initiale ;
6. la proposition en sortie des segments qui appartiennent au cluster choisi.

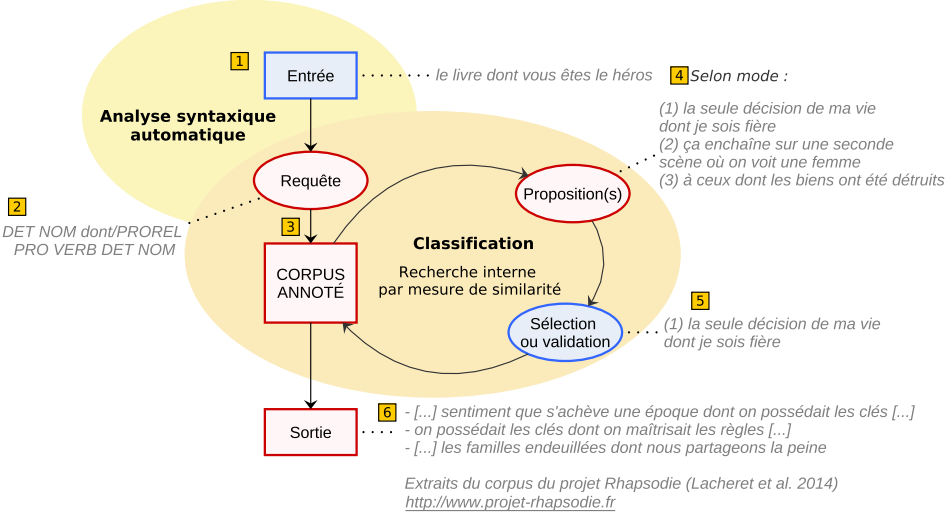

FIGURE 1 – Schéma du système de requêtes syntaxiques envisagé

S'agissant d'un projet en cours de développement, nous nous concentrons dans cette communication sur les trois premières étapes de notre chaîne.

3.2 La similarité comme méthode de recherche souple

Nous avons vu avec l'exemple des propositions relatives que la similarité syntaxique ne pouvait pas reposer uniquement sur une (suite d') unité(s) lexicale(s) mais devrait plutôt être décrite à l'aide de *patrons syntaxiques* sous la forme de séquences d'étiquettes, éventuellement mêlées à des mots. L'idée est bien entendu de pouvoir "matcher" des instances d'une même construction syntaxique en tolérant une certaine variation dans le lexique mais aussi dans la structure elle-même. En effet, si on regarde les propositions types de la Figure 1, on remarque que le premier segment donné par l'outil

3. Les étapes où l'utilisateur doit intervenir sont représentées par des items contournés de bleu et ont été réduites au strict nécessaire afin de rester en accord avec notre objectif de simplification.

ne correspond pas strictement à la requête. Pourtant, alors que leurs étiquettes diffèrent légèrement (on a respectivement "`[...] DET NOM PROREL PRO VERB ADJ`" pour la proposition et "`DET NOM PROREL PRO VERB DET NOM`" pour la requête), la proposition reste pertinente.

Puisqu'il n'est pas évident pour un utilisateur non-spécialiste de définir un patron efficace, c'est-à-dire avec un seuil de tolérance suffisamment élevé pour accepter les variations mais suffisamment bas pour ne pas manquer en précision, nous proposons une méthode basée sur la mesure d'une similarité entre le patron et les segments du corpus. Cette méthodologie a l'avantage d'être plus souple qu'une requête avec des expressions régulières et de permettre de rester plus proche des données en respectant l'approche bottom-up suggérée par le *Data-Driven Learning* [4]. Cette souplesse permet également à l'utilisateur d'avoir le choix entre plusieurs options :

1. faire une recherche en gardant le même mot ou certains mots de l'entrée (plutôt les mots grammaticaux : cas de la première proposition de l'étape 4 dans le schéma) ;
2. faire une recherche et récupérer des structures proches mais qui ne comportent pas forcément les mêmes mots (proposition 2, avec "où" au lieu de "dont") ;
3. si on dispose de ressources lexicales, faire une recherche avec un mot (cette fois-ci plutôt lexical) sémantiquement proche est également une possibilité.

La deuxième option correspond typiquement à la recherche de structures telles que les propositions relatives, puisqu'elles contiennent en français nécessairement une catégorie essentielle, celle du pronom relatif qui peut avoir différentes formes en surface parmi une liste finie. L'outil doit donc être capable d'identifier la catégorie des pronoms relatifs mais ne pas chercher forcément le même pronom que dans l'entrée, et surtout autoriser des variations dans les étiquettes périphériques étant donné que le contexte syntaxique peut être très différent selon la fonction du pronom dans la principale et la subordonnée. La première, quant à elle, se rapproche de ce que propose un concordancier, à la différence que le contexte doit être proche de celui de l'entrée, tandis que la troisième option intègrerait la possibilité d'étendre la requête en se servant de la similarité sémantique, comme c'est déjà le cas pour certaines applications en recherche d'information (moteurs de recherche, systèmes question-réponse) où le(s) mot(s)-clé(s) peuvent être remplacés par des synonymes ou hyperonymes.

Le choix entre ces différentes options pourrait être déterminé par l'utilisateur dès le départ s'il est suffisamment conscient de ce qu'il cherche et suffisamment compétent pour l'identifier. Dans le cas contraire, l'utilisateur pourra déterminer l'option qui lui convient le mieux grâce à la présentation d'un exemple similaire concret issu de chacune des options (voir étape 4 de la chaîne de traitements).

3.3 Mesures de similarité

Nous avons choisi d'utiliser les coefficients de Jaccard et de Dice, largement employées en TAL pour mesurer la similarité, en particulier entre deux mots ou deux chaînes de caractères. Dans notre contexte, il s'agit de mesurer la similarité entre des unités plus larges, des séquences d'étiquettes (`DET NOM PROREL...`) et/ou d'étiquettes couplées avec leur unité lexicale (dont/`PROREL`). Nous explorons également la piste de la distance d'édition (ou distance de Levenshtein), permettant d'évaluer indirectement une similarité. Si la similarité est maximale, la distance est nulle, et vice-versa. Cette alternative est particulièrement intéressante puisque la distance d'édition entre deux

4. Par opposition au *data-based* (littéralement "*basées sur les données*"), les méthodes dites *data-driven* (littéralement "*conduites par les données*") suivent un raisonnement inductif et partent de l'observation des régularités dans les données pour formuler des hypothèses ou les modifier.

"mots" (ou, similairement, entre deux séquences d'étiquettes) M et N se définit par le coût minimal nécessaire pour passer de M à N en effectuant des insertions, ajouts ou substitutions d'unités. Il est par ailleurs possible de pondérer le coût de chaque opération afin d'adapter le calcul de la distance d'édition à nos données. On pourrait en effet tout à fait considérer la suppresion d'un adjectif comme moins coûteux que la suppression d'un verbe ou d'une conjonction.

4 Expériences préliminaires

Des expérimentations sur le coréen langue étrangère sont actuellement réalisées, simulant notamment les recherches d'un apprenant qui éprouverait des difficultés à comprendre les contextes d'usage d'une structure grammaticale donnée et qui aurait besoin de davantage d'inputs authentiques en contexte (Wang, 2016). Nous donnons alors en entrée au programme des phrases qui sont typiquement à disposition des apprenants, celles qui servent à illustrer les explications grammaticales tirées de manuels de langue universitaires [5] et cherchons à comparer la structure de ces phrases à celles du Corpus Sejong (Kim *et al.*, 2007), le corpus de référence pour la langue coréenne. Les tests sont effectués sur le corpus annoté en partie du discours uniquement (environ 13,5 millions de tokens) et constitué d'échantillons de langue variés, écrits comme oraux.

5 Conclusion et perspectives

Nous avons vu qu'au coeur de notre étude se trouvait la simplification de l'accès aux corpus annotés pour un public non spécialiste, et bien que certaines études défendent que la confrontation à des données authentiques est bénéfique à un stade précoce de l'apprentissage (Holec, 1990; Boulton, 2009) la question de l'autonomie de l'apprenant face à la complexité des données authentiques se pose assurément, d'autant plus que nous avons fait le choix de travailler uniquement avec des corpus monolingues. Plusieurs options seront étudiées pour appréhender cette difficulté, à la fois au niveau de la présélection des données dès l'entrée (genre des textes et degré de lisibilité notamment) comme il est possible de le faire pour un grand nombre d'outils à des fins didactiques ou non, mais aussi au niveau de la visualisation des données en sortie (coloration syntaxique similaire à ce qui est proposé pour FipsColor (Nebhi *et al.*, 2010) ou encore dictionnaire intégré pour éviter que le vocabulaire n'ajoute une difficulté cognitive supplémentaire à l'analyse des résultats).

Un certain nombre de traitements sont envisagés sur le corpus, dont le regroupement préalable des phrases du corpus en clusters syntaxiques, améliorant ainsi la rapidité de l'outil puisqu'un seul membre représentatif de chaque cluster pourrait être comparé à la requête puis présenté à l'utilisateur. Cette étape supplémentaire évite à ce dernier d'être submergé de données non classées et devoir faire le tri parmi des dizaines voire centaines de résultats comme c'est souvent le cas avec un concordancier.

Notre outil n'est pas à proprement parler pédagogique en lui-même mais nous pensons que ce programme pourrait à terme compléter les ressources didactiques existantes en permettant une focalisation originale sur les structures grammaticales de la langue cible.

5. En l'occurrence, il s'agit des manuels des niveaux 1, 2 et 3 (équivalent à une à trois années d'études en coréen) de l'université de Yonsei et de ceux du centre linguistique de l'Université de Ewha.

Références

ANTHONY L. (2013). A critical look at software tools in corpus linguistics. *Linguistic Research*, **30**(2), 141–161.

AUGUSTINUS L., VANDEGHINSTE V. & VAN EYNDE F. (2012). Example-based treebank querying. In *Proceedings of eighth international conference on Language Resources and Evaluation (LREC'2012)*, p. 3161–3167.

BOULTON A. (2009). Testing the limits of data-driven learning : language proficiency and training. *ReCALL*, **21**(1), 37–54.

BOULTON A. (2012). Beyond concordancing : Multiple affordances of corpora in university language degrees. *Procedia-Social and Behavioral Sciences*, **34**, 33–38.

CAVALLA C. (2015). Collocations transdisciplinaire : réflexion pour l'enseignement. In *Le problème de l'emploi actif et / ou de connaissances passives des phrasèmes chez les apprenants de langues étrangères*. E.M.E & Intercommunication.

CHAMBERS A. (2005). Integrating corpus consultation in language studies. *Language learning & technology*, **9**(2), 111–125.

CHAMBERS A. (2010). L'apprentissage de l'écriture en langue seconde à l'aide d'un corpus spécialisé. *Revue française de linguistique appliquée*, **XV**, 9–20.

CHENG W., GREAVES C. & WARREN M. (2006). From n-gram to skipgram to concgram. *International journal of corpus linguistics*, **11**(4), 411–433.

FALAISE A., TUTIN A. & KRAIF O. (2011). Exploitation d'un corpus arboré pour non spécialistes par des requêtes guidées et des requêtes sémantiques. In *Actes de la 18e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2011)*, Montpellier, France.

HOLEC H. (1990). Des documents authentiques, pour quoi faire. *Mélanges Crapel*, **20**, 65–74.

JOHNS T. (1991). Should you be persuaded : Two samples of data-driven learning materials. *Classroom Concordancing : English Language Research Journal*, **4**, 1–16.

KIM H.-G., KANG B.-M. & HONG J. (2007). 21st Century Sejong Corpora (to be) Completed. *The Korean Language in America*, **12**, 31–42.

KRAIF O. & DIWERSY S. (2012). Le Lexicoscope : un outil pour l'étude de profls combinatoires et l'extraction de constructions lexico-syntaxiques. In *Actes de la 19e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2012)*, p. 399–406.

NEBHI K., GOLDMAN J.-P. & LAENZLINGER C. (2010). FipsColor : grammaire en couleur interactive pour l'apprentissage du français. In *Actes de la 17e conférence sur le Traitement Automatique des Langues Naturelles (TALN'2010)*, Montréal, Canada.

WANG I. (2016). A syntax-based query system adapted to language learning and teaching. In *American Association for Corpus Linguistics (AACL) and Technology for Second Language Learning (TSLL) Conference*, Ames, USA : Iowa State University. Poster presentation.