Comparaison de la validité de différentes mesures statistiques pour l'extraction de collocations fondamentales dans un corpus issu du Web

Veronica Benigno, Olivier Kraif

Comparaison de la validité de différentes mesures statistiques pour l'extraction de collocations fondamentales dans un corpus issu du Web

Veronica Benigno[1], Olivier Kraif[2]
Univ. Grenoble Alpes, LIDILEM, F-38040 Grenoble [1, 2]

L'étude présentée ici est tirée d'un travail de recherche plus vaste, une thèse de doctorat portant sur le concept de « collocation fondamentale », étudié à partir de l'analyse d'un corpus de grande dimension tiré du Web (Benigno, 2011). L'apprentissage du vocabulaire d'une langue seconde est un processus d'acquisition progressif dont la première étape est la maîtrise du *vocabulaire fondamental*, comprenant les unités les plus simples et les plus fréquentes, réputées pour leur utilité dans les actes communicatifs de la vie quotidienne. De Henmon (1924) à Gougenheim (1954) pour le français, de Thorndike (1929) à West (1950) pour l'anglais, de nombreux auteurs ont très tôt cherché à définir, à caractériser et à énumérer ces unités formant le socle de l'acquisition du lexique. Mais les unités fondamentales ne forment pas une liste de mots isolés : ces mots s'inscrivent dans des ensembles plus larges, que l'on parle d'expressions préfabriquées, de combinaisons typiques ou de simples collocations. Dans cette perspective, on peut définir les *collocations fondamentales* comme des unités significatives, fondées sur une relation comportant un certain degré de figement entre une base et un collocatif, fréquentes dans l'usage et conceptuellement disponibles (Gougenheim, 1958). Ces unités, qui représentent les co-occurrences les plus élémentaires d'un mot simple donné, correspondent aux réalisations les plus fréquentes, en contexte, de ses différents sens. De très nombreux travaux se sont attachés à la définition du concept de collocation (Hausmann, 1979, Mel'cuk, 1984, Williams, 2003, Grossmann et Tutin 2003), celui-ci pouvant s'appliquer à des unités allant de la combinaison libre aux expressions idiomatiques totalement figées, mais l'intérêt s'est plus souvent porté sur les langues de spécialité que sur la langue générale ; en outre, bien que ces phénomènes aient été largement traités dans les recherches sur l'enseignement d'une langue seconde, peu d'études se sont intéressées à la notion de collocation fondamentale.

Un des objectifs de la thèse de Benigno (2011) était d'évaluer l'importance relative de la fréquence dans la définition de ce concept, par rapport à ses autres fonctions. Cela nous a amené à utiliser différents indicateurs statistiques, dont nous effectuons ici une comparaison, en fonction des spécificités de notre corpus.

Afin d'observer la fréquence des collocations fondamentales d'une manière fiable, nous avons dû extraire des événements statistiquement significatifs pour les cooccurrences (et non seulement les occurrences prises isolément), ce qui nous a contraint à utiliser un corpus de très grande taille. La construction d'un corpus de langue générale de grand volume est un objectif très ambitieux, hors de la portée de nos moyens. Il nous est apparu que le corpus *frWack*, partie française du corpus *Wacky* (Baroni et al., 2009), répondait de manière satisfaisante à nos besoins spécifiques : comportant plus d'un milliard de mots, il est composé de textes qui représentent une grande variété de genres, de types et de domaines. Mais comme tout corpus web récupéré automatiquement, *frWac* présente des inconvénients. D'une part, le corpus n'est pas conçu pour rendre possible un échantillonnage, afin de construire un sous-corpus équilibré, en tenant compte de la proportion relative des genres textuels et des domaines caractéristiques de la langue générale. D'autre part il est susceptible de contenir des données bruitées, puisque toutes les informations publiées sur le Web ne peuvent pas être considérées comme du texte, d'un point de vue traditionnel. Les pages Web contiennent en effet des textes "*boilerplate*" (menu, pieds de page, mentions légales, etc), des fragments de texte, des citations répétées, des listes, des tableaux, etc. Baroni et al. (2009) expliquent quels traitements ont été appliqués pour nettoyer automatiquement le frWack de toutes ces données non-textuelles (ce qui est indispensable, car celles-ci sont susceptibles de brouiller les statistiques globales), ainsi que des redondances. Mais il faut être conscient que malgré l'efficacité des méthodes employées, tout le

bruit n'a pas été éliminé, et certaines répétitions peuvent être liées aux redondances du Web (un même texte pouvant être publié sur de très nombreux sites, par copier/coller) plus qu'aux répétitions inhérentes à l'usage « normal » de la langue (quoique la frontière ne puisse être tracée de façon définitive entre ces deux types de redondances, naturellement). Le corpus est livré gratuitement en format XML, et comporte un découpage en texte (avec l'url source), en phrases, et pour chaque phrase, les sorties de l'étiqueteur treetagger (Schmid, 1995) sur trois colonnes (forme, catégorie et lemme).

Pour exploiter ce corpus gigantesque, nous avons développé des scripts Perl ad hoc, nous permettant d'extraire rapidement les cooccurrents les plus fréquents pour un pivot donné, accompagné des mesures d'association standard basées sur les fréquences d'occurrence et de co-occurrence (t-score, z-score, loglike et PMI), étudiées notamment par Evert (2008). Un autre indicateur calculé grâce à ces scripts est la *dispersion*, mesurant le nombre de noms de domaine différents dans lesquels une collocation a été observée. Cette mesure permet en quelque sorte de contrôler le caractère général d'une collocation, afin de mettre de côté des cooccurrences fréquentes qui n'apparaîtraient que sur un petit nombre de sites spécialisés. Outre ces statistiques, nos scripts extraient les concordances liées aux différents couples base - collocatif, afin d'étudier les collocations en contexte.

Enfin, parallèlement à ces observations quantitatives, nous avons demandé à des locuteurs natifs de juger du caractère "fondamental" des collocations extraites. En effet, d'après Stubbs (2002) la définition du vocabulaire de base ne dépend pas seulement de la fréquence, mais aussi de critères fonctionnels tel que ce qui est pertinent dans la communication pour des enfants ou des locuteurs non-natifs, et ce qui est jugé comme plus important à enseigner pour des apprenants débutants en langue seconde.

Les résultats obtenus montrent qu'il existe une corrélation significative quoique non systématique entre la fréquence observée et le jugement des locuteurs - certains points singuliers montrant que la disponibilité et le degré de figement sont bien des critères complémentaires permettant de distinguer certaines collocations moins fréquentes.
Autre observation assez inattendue : la fréquence brute et le z-score donnent ici les meilleurs résultats. L'information mutuelle, le t-score le loglike sélectionnent essentiellement des co-occurrents fréquents mais sans généralité, dont la dispersion est faible, et qui correspondent le plus souvent à des entités nommées ou à du bruit. Ce fait est étroitement lié à la structure du corpus, et notamment à l'étroitesse de la période temporelle du moissonnage (10 jours), ce qui est source de biais étant donnée la sur-représentation de certains thèmes liés à l'actualité (de nombreux contenus trouvés sur le Web étant liés aux événements prochains et récents). Nous notons que le loglike pourrait donner de très bons résultats (ce qui est généralement le cas pour l'extraction des collocations) s'il était combiné à un seuil minimum de dispersion.
Au final, ces observations nous permettent de suggérer quelques pistes d'amélioration pour l'extraction automatique de collocations fondamentales à partir des données du Web.

Références

Baroni M., Bernardini S., Ferraresi A., Zanchetta E. (2009) The WaCky Wide Web : A Collection of Very Large Linguistically Processed Web-Crawled Corpora. *Language Resources and Evaluation*, 43 (3), p. 209-226.

Benigno V. (2012) *La notion de collocation fondamentale. Etude de corpus en vue d'une exploitation didactique*. Thèse de doctorat en cotutelle, Université de Grenoble et Université de Palerme.

Evert, S. (2008) Corpora and collocations. In A. Lüdeling and M. Kytö (eds.), *Corpus Linguistics. An International Handbook*, article 58. Mouton de Gruyter, Berlin.

Gougenheim G. (1958/1971) *Dictionnaire fondamental de la langue française*. Didier Edition internationale.

Grossmann, F., & Tutin, A. (Ed.), 2003) *Les collocations : analyse et traitement*. Amsterdam, De Werelt.

Hausmann F. J. (1979) Un dictionnaire des collocations est-il possible ? *Travaux de linguistique et de littérature* XVII (1), Strasbourg, p. 187-195.

Henmon V. A. C. (1924) *A French Word Book*. University of Wisconsin, Bureau of Educational Research, Bulletin n 3.

Mel'čuk I. et al. (1984/1988) *Dictionnaire explicatif et combinatoire du français contemporain*. Montréal, Canada, Presses de l'Université de Montréal.

Péchoin D. (1991) *Thésaurus Larousse, Des mots aux idées, des idées aux mots*. Paris, Larousse.

Sinclair J. (1991) *Corpus, concordances, collocation*. Oxford, Oxford University Press.

Schmid H. (1995): Improvements in Part-of-Speech Tagging with an Application to German. *Proceedings of the ACL SIGDAT-Workshop*. Dublin, Ireland.

Stubbs M. (2002) *Words and phrases : corpus studies of lexical semantics*. Oxford, Blackwell Publishing.

Thorndike E. L. (1921) *The Teacher's Word Book*. New York Teachers College, Columbia Univerity.

West M. (1953) *A General Service List of English Words*, London, Longman.

Williams G. (2003) « Les collocations et l'école contextualiste britannique », dans Grossmann F., Tutin A. (Ed.), *Les collocations : analyse et traitement*. Amsterdam, de Werelt, p. 33-44.