

Fonctions lexicales et annotation de corpus

Des outils pour l'apprentissage des collocations linguistiques

Isabella Carini, Adriano Ferraresi, Silvia Bernardini (1),
François Lareau (2)

(1) Dipartimento di Interpretazione e Traduzione (DIT), Università di Bologna

(2) Observatoire de linguistique Sens-Texte (OLST), Université de Montréal

Résumé : Cette étude présente une méthodologie pour l'annotation par fonctions lexicales des sous-corpus italiens du *European Parliament Translation and Interpreting Corpus* (EPTIC). Nous proposons une méthode pour étiqueter certaines collocations selon les catégories sémantiques fournies par les fonctions lexicales de la Théorie Sens-Texte (TST). Le corpus ainsi annoté deviendrait alors une ressource utile pour l'apprentissage des collocations de la langue italienne, en particulier pour les traducteurs et interprètes non italophones.

Mots-clés : collocations, fonctions lexicales, Théorie Sens-Texte, annotation de corpus

1 Introduction : pourquoi un corpus annoté ?

Les collocations jouent un rôle clé dans l'apprentissage des langues : leur maîtrise est un élément essentiel d'une compétence linguistique élevée, mais très difficile à atteindre.

Depuis une vingtaine d'années, l'étude des expressions polylexicales a reçu une attention croissante en didactique des langues, grâce à l'élaboration de nouvelles approches didactiques centrées sur la phraséologie (Nattinger et DeCarrico, 1992 ; Nesselhauf, 2005), et en lexicographie, avec la parution de ressources spécialisées comme les dictionnaires de collocations. Tout cela semble suggérer qu'aujourd'hui, l'apprentissage des collocations est plus aisé qu'il ne l'était auparavant, quand la phraséologie ne faisait pas encore l'objet d'études spécialisées. Et pourtant, surtout dans la forma-

tion d'interprètes et de traducteurs, l'apprentissage des collocations reste un point critique (Primož, 2011), et on remarque que les ressources disponibles n'ont pas complètement résolu des difficultés de fond. Depuis toujours, la nature idiosyncratique et imprévisible du lien entre les éléments d'une collocation (la base et ses collocatifs) pose des difficultés même aux étudiants de niveau avancé et rend très compliqué leur catégorisation et leur enseignement : on a du mal à enseigner le lexique et la phraséologie de façon aussi systématique que la grammaire. De plus, les ressources lexicographiques disponibles pour l'apprentissage des collocations présentent souvent des limites importantes.

1.1 Les limites des dictionnaires de collocations

Les dictionnaires de collocations de l'italien actuellement disponibles sont peu nombreux : à notre connaissance, il n'y a que le *Dizionario delle collocazioni* (Tiberii, 2012), le *Dizionario delle combinazioni lessicali* (Urzi, 2009), et le *Dizionario Combinatorio Italiano* (Lo Cascio, 2013). Le *CombiNet* mérite une mention à part : il s'agit d'un « corpus-based online dictionary of Italian word combinations » (Castagnoli *et al.*, 2015), un projet de l'Université Roma 3, de l'Université de Pise et de l'Université de Bologne, qui malheureusement n'est pas encore disponible pour consultation. Ici nous nous limiterons à esquisser les caractéristiques principales de ces ressources, car une analyse approfondie et exhaustive des dictionnaires de collocations de l'italien dépasse l'objectif de notre étude.

Le dictionnaire de Tiberii (2012) partage la même faiblesse que Smadja (1991) et Fontenelle (1992, 1998) avaient identifié il y a vingt ans dans la plupart des dictionnaires de collocations : les collocatifs sont regroupés tout simplement par partie du discours, et en ordre alphabétique, sans aucune distinction sémantique. Cela rend la consultation particulièrement frustrante pour un usager non italophone, à qui le dictionnaire de Tiberii ne fournit aucun repère pour s'orienter et choisir le bon collocatif pour exprimer un sens donné. Le dictionnaire de Urzi (2009) présente lui aussi une macro-division des collocatifs selon leur partie du discours, mais, dans chaque sous-section, Urzi regroupe les collocatifs par aire sémantique (les collocatifs verbaux dont les sens se ressemblent sont listés ensemble, et ils sont séparés des autres verbes par un |). Malgré cela, ces aires sémantiques ne sont pas étiquetées de façon explicite, et les différences entre les collocatifs qui partagent la même aire ne sont pas signalées. Le dictionnaire de Urzi fournit plus d'information : il signale les collocatifs qui constituent des antonymes l'un de l'autre, et il offre des exemples d'usage qui ont une valeur de définition, même si ces exemples ne sont pas systématiques dans le dictionnaire. Globalement, le dictionnaire de Urzi est un outil pour des usagers italophones (qui constituent les destinataires principaux de cet ouvrage, selon l'introduction) ou des étudiants de niveau avancé, mais pour des usagers qui ont une faible connaissance passive de l'italien sa consultation reste compliquée à cause de l'absence d'information sémantique, qui rend nécessaire le recours à d'autres ressources (notamment des dictionnaires traditionnels) pour déchiffrer le sens des collocatifs. C'est précisément sur ce point que le dictionnaire de Lo Cascio (2013), spécialement dans sa version numérique en ligne, se révèle la ressource la plus complète et utile, car en plus de la division classique par partie du discours et ses nombreux exemples d'usage, chaque

collocatif est lié à sa définition. En plus, pour certaines collocations, on trouve des gloses explicatives entre crochets qui aident l'utilisateur à s'orienter entre les différents sens.

Malgré leur richesse, le problème principal des dictionnaires de collocations est que, dans ces ressources, les collocatifs fréquents et rares sont mis au même plan, et toute indication sur le degré d'association des collocatifs — tout comme sur leur emploi en domaines ou typologies textuelles spécifiques — est absente. Pour un étudiant non-italophone, il serait par conséquent très compliqué de comprendre quels collocatifs sont plus communs ou plus adaptés selon les différents contextes.

1.2 Les corpus comme outils d'apprentissage

Pour ce qui concerne ce dernier aspect, c'est-à-dire la possibilité d'observer les collocatifs en contexte, leur fréquence et leur distribution, les corpus peuvent aider : l'utilité des corpus en didactique des langues et dans la formation de traducteurs et d'interprètes est désormais reconnue (Bernardini *et al.*, 2003 ; Sinclair, 2004). Grâce à l'introduction des corpus dans les parcours universitaires pour interprètes et traducteurs, les étudiants disposent d'outils comme les concordanciers pour mener de façon autonome des recherches sur les collocations d'une lexie donnée, obtenant ainsi des collocatifs « réels » (et non pas le fruit de la sélection d'un auteur), avec des données sur leur fréquence. De plus, l'approche par corpus offre la possibilité de raffiner les critères de la recherche : en sélectionnant un corpus de langue générale, on pourra observer les fréquences générales de tel ou tel collocatif, tandis qu'en limitant le champ à des corpus de spécialité, il est possible d'observer quels collocatifs apparaissent le plus souvent dans des genres textuels ou des domaines spécifiques. Si les corpus peuvent aider les étudiants à s'orienter entre les collocatifs, le problème de l'absence de distinctions sémantiques demeure : un corpus, s'il n'est pas doté d'une annotation sémantique, ne fournit aucune information sur le lien entre les éléments d'une collocation, ni sur les différences de sens entre les différents collocatifs. L'étudiant allophone est obligé, encore une fois, de consulter un dictionnaire de langue.

Pour toutes ces raisons, nous avons reconnu la nécessité d'un nouvel outil qui faciliterait le repérage, l'étude et l'apprentissage des collocations. L'outil que nous proposons consiste en une annotation des collocations en corpus basée sur les fonctions lexicales de la théorie Sens-Texte (Mel'čuk, 1996).

1.3 Notre proposition : une annotation par fonctions lexicales

Les fonctions lexicales (FL) de la Théorie Sens-Texte (TST) constituent une notation formelle permettant de représenter et de modéliser de façon systématique les collocations (en particulier les liens sémantiques et syntaxiques entre la base et le collocatif), et de classer les collocatifs d'une lexie selon un critère sémantique.

Dans le cadre de la TST, on définit une collocation non seulement en termes de cooccurrence préférentielle entre mots, mais plus précisément comme une expression formée par plusieurs éléments qui entretiennent une relation lexicale stable, où un des éléments — la base — est choisi librement par le locuteur pour exprimer un

sens donné, tandis que le choix d'un autre élément — le collocatif — est sujet à des restrictions en fonction de la base (Mel'čuk, 1998)

Prenons ces collocations de l'italien :

stanco morto (très fatigué)

ricco sfondato (très riche)

povero in canna (très pauvre)

Dans les trois cas, la *base* de la collocation (*stanco*, *ricco*, *povero*) est accompagnée par un collocatif qui exprime le sens 'très'. Le lien entre base et collocatif possède un degré de figement et de spécificité : chaque base a un collocatif différent pour exprimer le sens 'très' (*morto* ne représente le sens d'intensité qu'en relation à *stanco*, on ne peut pas dire « *ricco morto* »), donc les collocatifs varient selon la base, mais dans ces exemples il existe un rapport sémantique constant entre chaque base et son collocatif.

À l'aide de la notation formelle de Mel'čuk, on peut représenter ce rapport constant comme une fonction mathématique $f(x) = y$, en appelant Magn la fonction qui représente le lien entre base et collocatif : les collocations de notre exemple deviennent ainsi trois instances de la même fonction lexicale $\text{Magn}(x) = y$:

$\text{Magn}(\text{stanco}) = \text{morto}$

$\text{Magn}(\text{povero}) = \text{in canna}$

$\text{Magn}(\text{ricco}) = \text{sfondato}$

Cette notation formelle n'est pas facile à approcher, mais comme Polguère (2003) le remarque, les fonctions lexicales peuvent être appréhendées sans qu'il soit nécessaire de recevoir une formation véritable en linguistique Sens-Texte, à travers leur vulgarisation. Dans le cadre de cette étude, comme on le verra, nos instruments de vulgarisation sont des gloses insérées dans l'annotation de corpus.

Les FL constituent donc un outil à exploiter pour l'apprentissage des collocations (Ancil, 2005 ; Mel'čuk et Polguère, 2007 ; Tremblay, 2009 ; Tremblay et Polguère, 2014 ; Frassi et Tremblay, 2016).

L'annotation proposée ici consiste à assigner à chaque collocatif d'une lexie une étiquette avec la FL qui lui correspond. De cette façon, deux types d'informations seraient explicitées : information syntaxique (par exemple, la configuration des actants syntaxiques profonds d'un collocatif verbal) et information sémantique. Grâce à cette annotation, il serait enfin possible de distinguer les collocatifs par leur sens.

L'idée à la base de notre travail a déjà été exploitée (Tutin, 2005 ; Augustyn et Tutin, 2009), mais ici nous l'appliquons à une langue différente (l'italien) et avec une méthode différente, comme nous le verrons.

Le corpus ainsi annoté constituerait une ressource qui présenterait plusieurs avantages :

1. Il permet l'observation des collocations en contexte, ce qui facilite la mémorisation et la réutilisation des expressions (Nesselhauf, 2005) ;
2. Il permet d'observer et de mémoriser les propriétés syntaxiques des collocations, qui doivent être apprises en même temps que la collocation elle-même, par exemple le type d'article (déf/indéf ; sing/plur) qui est requis dans une collocation, ou la voix d'un verbe : on peut observer que certains verbes en collocation se trouvent bien plus fréquemment à la voix passive qu'à la voix active (par exemple *être saisi* d'angoisse est plus fréquent que *l'angoisse saisit*). Souvent les dictionnaires ne fournissent pas cette information, qui en revanche est très bien observable en corpus (Augustyn et Tutin, 2009) ;
3. Il permet d'observer la fréquence et la distribution des collocations (certaines collocations sont plus typiques d'un domaine, d'une typologie textuelle, d'un niveau de langue donné).

2 Méthodologie

2.1 Le corpus choisi : EPTIC

Le corpus choisi pour l'annotation est EPTIC (*European Parliament Translation and Interpreting Corpus*), un corpus constitué des transcriptions des interventions au Parlement européen (et de leurs traductions et interprétations) en anglais, français et italien. Il s'agit d'un corpus trilingue, parallèle et comparable à la fois, réalisé à partir de 2013 au Département d'Interprétation et traduction de l'Université de Bologne (Ferraresi et Bernardini, sous presse), sous la direction Silvia Bernardini, Adriano Ferraresi et Maja Miličević Petrović. Le choix du corpus EPTIC pour cette étude est motivé par deux facteurs principaux : 1) les textes institutionnels européens sont une typologie textuelle très exploitée dans la formation des traducteurs et interprètes, d'où l'intérêt d'utiliser ce corpus comme base pour un outil destiné à l'apprentissage ; 2) notre annotation implique une série d'opérations à effectuer manuellement, donc la dimension raisonnable du corpus (330 000 mots environ) facilite la tâche.

Notre étude consiste en une proposition méthodologique : pour cette raison, l'annotation n'a pas été réalisée sur la totalité des collocations du corpus. Nous avons choisi de nous concentrer sur une seule langue (l'italien, c'est-à-dire les quatre sous-corpus italiens d'EPTIC, 70 000 mots environ) et sur un seul lexème (DIRITTO 'droit'). Notre étude de cas vise à présenter les bénéfices potentiels de ce type d'annotation si elle était réalisée sur l'ensemble des lexèmes d'un corpus (et, idéalement, sur un corpus de dimension plus importante).

La raison pour laquelle nous avons choisi le lexème DIRITTO est qu'il s'agit d'un substantif relativement fréquent dans les sous-corpus italiens d'EPTIC, et qu'il présente des collocatifs intéressants aux fins de notre annotation. La sélection du lexème DIRITTO a été réalisée selon les critères et les étapes que nous présentons ci-dessous.

2.2 Critères de sélection du matériel linguistique à annoter

Premièrement, nous avons défini deux critères de sélection générale, afin d'isoler un échantillon intéressant :

- la base de la collocation doit nécessairement être un lexème (un mot sémantiquement plein) et plus précisément un substantif, parce que la plupart des FL correspondent à des collocations qui ont pour base un substantif (Mel'čuk, 1998) ;
- pour travailler sur un échantillon significatif, la base de la collocation doit être relativement fréquente dans les quatre sous-corpus. Selon Sinclair (2005) « Un mot qui n'est pas spécialement ambigu requiert au moins vingt instances pour permettre à des lexicographes d'esquisser une description de son comportement. ». Toutefois, en raison de la dimension limitée du corpus EPTIC, nous avons décidé de fixer deux seuils arbitraires de fréquence minimale : ≥ 10 occurrences dans au moins un des sous-corpus italiens, et ≥ 40 occurrences en total.

Ainsi nous avons isolé les substantifs les plus fréquents dans les sous-corpus italiens (Table 1).

TABEAU 1 : Lexèmes issus de la première sélection de matériel linguistique.

Substantif	Occurrences totales
COMMISSIONE	320
UNIONE	310
PRESIDENTE	302
ACCORDO	269
DIRITTO	179
PARTE	145
MERCATO	106
PROBLEMA	97
OBIETTIVO	67
IMPEGNO	61

Ensuite, nous avons observé les concordances de chacun des substantifs précédemment sélectionnés et avons opéré une seconde sélection sur la base des critères qualitatifs suivants :

- variété de contextes : chaque substantif doit présenter une certaine variété de collocatifs, pour représenter une bonne variété de FL. Pour cette raison, nous avons éliminé les substantifs qui apparaissent dans des contextes répétitifs, qui ne sont pas intéressants aux fins de l'étude, par exemple le lexème UNIONE (259 cooccurrences avec EUROPEA sur 320 occurrences totales, soit 80,93%) et PRESIDENTE (dans 39% des cas il se trouve en cooccurrence avec SIGNOR) ;
- présence de collocatifs verbaux : les collocatifs verbaux correspondent souvent à des FL standards, par exemple les verbes supports ou de réalisation (Mel'čuk, 1998). Pour cette raison, nous avons sélectionné en priorité les substantifs dont au moins 50 % des collocatifs sont des verbes ;
- présence de collocations au sens décrit par Mel'čuk, c'est-à-dire expressions formées par plusieurs éléments qui entretiennent une relation lexicale stable (cf. 1.3). Nous avons donc rejeté les substantifs dont les cooccurrences avec

d'autres lexèmes ne sont pas des collocations, mais plutôt des associations libres entre mots ;

- correspondance avec des FL standards : les collocations de notre corpus, afin de pouvoir être annotées, doivent être représentables avec des FL standards ; tous les substantifs dont la plupart des collocatifs correspondent à des fonctions non standard ont été rejetés.

Après cette deuxième sélection, nous avons finalement choisi le lexème DIRITTO, qui satisfait tous les critères ci-dessus. Il est nécessaire de préciser que le lexème DIRITTO correspond à deux sens dans notre corpus : 'ensemble de normes juridiques' (*le droit international*) et 'faculté' (*les droits de l'homme*). Après cette désambiguïsation, qui a montré 26 occurrences du premier sens et 153 du deuxième sens, nous avons choisi ce dernier, dont 73 des 153 occurrences correspondent à des FL standards.

L'extraction des termes aurait pu se faire à l'aide de méthodes semi-automatiques (Evert, 2008) ou des outils comme SketchEngine, mais vu la taille raisonnable de données à traiter et les visées pédagogiques de cette étude (qui est née en tant que mémoire de maîtrise), nous avons opté pour un traitement manuel.

2.3 Références théoriques pour l'annotation

Toute annotation linguistique qui se veut rigoureuse doit se référer à une approche théorique définie et possiblement à des descriptions déjà existantes (Leech, 2005). Dans notre cas, l'approche choisie est la TST, et idéalement notre annotation devrait se référer à un dictionnaire ou à une base de données où les collocations de l'italien sont classées à l'aide du modèle des FL. Comme nous l'avons déjà mentionné, notre travail est proche de celui d'Augustyn et Tutin (2009), qui utilisaient comme référence théorique une base de données de collocations développée par Fontenelle (1997), qui codifie les collocations en exploitant les catégories fournies par les FL. Malheureusement, pour l'italien il n'existe pas de telle base de données, ni de ressources lexicographiques inspirées de la TST¹. Notre première tâche a donc été l'analyse des données lexicales et leur catégorisation : nous avons rédigé une fiche lexicographique du lexème DIRITTO, en nous appuyant sur les critères du *Dictionnaire explicatif et combinatoire du français contemporain* (DECFC) (Mel'čuk *et al.*, 1984-1999) et en utilisant comme référence le DiCoLiLex, une base de données électronique de collocations du français actuellement en cours de réalisation à l'Université de Montréal (Marengo et Robichaud, 2017). Ces instruments ont été utiles en partie, car bien que le français et l'italien soient des langues typologiquement proches, d'importantes différences entre elles ne permettent pas d'appliquer sur l'italien la catégorisation du français telle quelle. Ainsi, pour la réalisation de notre fiche lexicographique, un travail d'analyse sémantique, syntaxique et lexicale du lexème DIRITTO a été nécessaire. Notre fiche, suivant le modèle du DECFC, contient de l'information sur la structure des actants sémantiques et syntaxiques profonds du lexème DIRITTO, et une liste de FL avec

1. Les ressources lexicographiques de ce type pour la langue française sont nombreuses, à partir du Dictionnaire Explicatif et Combinatoire (Mel'čuk *et al.*, 1984-1999), en passant par le Lexique Actif du Français (Mel'čuk et Polguère, 2007), jusqu'aux ressources de la famille DiCo de l'Université de Montréal, le Réseau Lexical du Français (Lux-Pogodalla et Polguère, 2011) et le projet ORTOLANG (Open Resources and TOols for LANGuage, www.ortolang.fr).

les collocatifs qui correspondent à chaque fonction. La fiche ainsi élaborée (*cf.* Carini, 2018) a été notre référence pour l'annotation.

2.4 Schéma d'annotation et codage

Les étiquettes d'annotation ont été conçues selon les critères de Leech (2005), qui souligne l'importance d'utiliser des étiquettes *brèves, univoques et transparentes pour l'utilisateur* dans l'annotation de corpus. Nos étiquettes correspondent aux noms des différentes FL, qui possèdent les caractéristiques citées par Leech. La liste des étiquettes utilisées, avec les collocatifs auxquels elles correspondent et leur fréquence dans les sous-corpus italiens de EPTIC est donnée à la Table 2. Chaque collocatif est étiqueté de cette façon : nom de la FL, suivi de la base, suivi d'un numéro unique (ex. : Real1*diritto*71, voir Figure 1). Il faut préciser que la base de chaque collocation est également annotée avec une étiquette « base* » suivie d'un numéro unique qui permet de la lier à son collocatif (voir Figure 1).

Le corpus EPTIC possède deux niveaux d'annotation de base : une annotation par partie du discours et une lemmatisation. EPTIC a été précédemment annoté avec TreeTagger², dont le format de codage est un format vertical *one word per line CWB-compliant* (Ferraresi et Bernardini, sous presse). Notre annotation a donc été codée en format vertical, en ajoutant deux colonnes supplémentaires à la structure du corpus.

L'annotation a été réalisée manuellement. L'automatisation du processus n'a pas été possible à cause de deux facteurs principaux : l'absence d'une ressource comme celle utilisée par Augustyn et Tutin (2009), soit une base de données lexicales où les collocations sont déjà codées en termes de FL, et l'absence d'annotation sémantique dans le corpus EPTIC, qui a entraîné la nécessité de désambiguïser manuellement les occurrences de DIRITTO en tant que 'faculté, prérogative' et en tant qu'ensemble de normes juridiques'.

Les étiquettes de notre annotation permettent, entre autres, de distinguer :

- les verbes qui ont un sens de réalisation par rapport au lexème DIRITTO ('faire ce qu'on est censé faire avec un droit', c'est-à-dire le respecter). Ces verbes sont étiquetés avec la FL Real. En combinaison avec Anti, le sens est inversé : la FL AntiReal indique les verbes qui expriment le sens 'faire le contraire de ce qu'on est censé faire avec un droit' (ne pas le respecter) ;
- les verbes qui expriment le sens de causation, en particulier le sens 'faire en sorte que un droit continue d'exister'. Ces verbes sont étiquetés avec la fonction PermFunc ;
- les noms dérivés de ces verbes : ces noms sont étiquetés par une fonction lexicale complexe qui commence par S₀ (fonction lexicale qui représente la substantivation).

2. <http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/> (consulté le 01.12.2017).

TABLEAU 2 : Collocatifs de DIRITTO extraits des sous-corpus italiens de EPTIC, leurs étiquettes et nombre d'occurrences.

Collocatif	Étiquette (FL)	Occurrences
<i>Avere</i>	Oper ₁	6
<i>Beneficiare</i>	Real ₁	4
<i>Garantire</i>	Real ₃	5
<i>Riconoscere a X [un diritto]</i>	Real ₄₁	3
<i>Negare a X [un diritto]</i>	AntiReal ₄₁	3
<i>Riconoscere</i>	Real ^I ₄	4
<i>Rispettare</i>	Real ^{II} ₄	8
<i>Violare</i>	AntiReal ^{II} ₄	1
<i>Intaccare</i>	AntiReal ^{III} ₄	2
<i>Rispetto</i>	S. Real ^{II} _{0 4}	3
<i>Nel rispetto</i>	Adv. Real ^{II} _{0 4}	3
<i>Violazione</i>	S. AntiReal ^{II} _{0 4}	2
<i>Difendere</i>	PermFunc ₀	4
<i>Proteggere</i>	PermFunc ₀	2
<i>Tutelare</i>	PermFunc ₀	12
<i>Difesa</i>	S. PermFunc _{0 0}	2
<i>Tutela</i>	S. PermFunc _{0 0}	6
<i>Difensore</i>	S. PermFunc _{1 0}	3

Il est évident que, afin de pouvoir exploiter efficacement les informations contenues dans l'annotation, l'utilisateur devrait posséder des notions de base relatives aux FL. Pour cette raison notre annotation prévoit l'ajout de deux colonnes a la structure du corpus (voir Figure 1) :

	POS	lemma	LexF	LF_glose
i	ART	il		
pazienti	NOUN	paziente		
potranno	VER2:fin	potere		
beneficiare	VER:infi	beneficiare	Real1_diritto_71	Fare_uso,realizzare_lo_scopo
del	ARTPRE	del		
diritto	NOUN	diritto	base_71	
alla	ARTPRE	alla		
mobilità	NOUN	mobilità		

FIGURE 1: Exemple du format vertical de l'annotation.

- **LexF** pour les fonctions lexicales
- **LF_glose** pour la glose de chaque FL, qui sert à expliquer vulgariser la notation formelle, sur le modèle des gloses utilisées dans le *Lexique Actif du Français* (Mel'čuk et Polguère, 2007) et dans Augustyn et Tutin (2009). On estime que, avec le support des gloses, l'utilisateur pourrait se servir de l'annotation sans connaître en détail le système des FL.

3 Résultats attendus et difficultés

3.1 Mode de consultation

Notre corpus annoté, qui dans le cadre de notre proposition méthodologique a une pure valeur d'exemple, n'a pas été indexé, car l'annotation n'a pas été réalisée sur l'ensemble des lexèmes du corpus. Si le corpus annoté était indexé, il pourrait être consulté à travers NoSketch Engine³. Différents modes d'exploration seraient alors possibles. On pourrait choisir d'observer les concordances du lexème DIRITTO, où les collocatifs apparaîtraient avec l'étiquette de la fonction lexicale correspondante et la glose, comme illustré en Figure 2, où la fonction explicative de la glose est bien visible : le verbe *tutelare* est un collocatif de DIRITTO auquel correspond la FL PermFunc0 ; en soi, cette FL est assez opaque pour un usager qui n'est pas familier avec le modèle, mais la glose vient expliciter le sens de la FL : « faire en sorte que quelque chose continue à exister ».

ogni	ADJ	paese	NOM	detiene	VER pres /Oper1_diritto_4 /avere	il	DET det	diritto	NOM /base_4	a	PRE	sceglie	VER inf	la	DET
Dobbiamo	VER pres	tutelare	VER inf /PermFunc0_diritto_34 /fare in modo che continui a esistere	il	DET det	diritto	NOM /base_34	del	PRE det	consumatore	NOM				
nega	VER pres /AntiReal41_diritto_60 /fare il contrario di ciò che si deve	alle	PRE det	comunità	NOM	il	DET det	diritto	NOM /base_60	di	PRE	vivere	VER inf		

FIGURE 2: Concept de visualisation des concordances avec NoSketchEngine.

On pourrait également effectuer des recherches ciblées à l'aide d'expressions régulières (voir Table 3).

3. (consulté le 11.02.2018).

TABLEAU 3 : Exemple d'interrogations sous forme d'expressions régulières.

Recherche désirée	Expression régulière (RegEx)
Afficher toutes les occurrences de DIRITTO qui sont associées à des collocatifs correspondant à des FL	[LexF="base_.*"]
Trier les collocatifs selon leur partie du discours:	[LexF="S._diritto_.*"]
Afficher tous les collocatifs nominaux de DIRITTO	
Afficher tous les collocatifs qui correspondent à une FL donnée (ex :Oper1)	[LexF="Oper1_diritto_.*"]

Le mode d'exploration qui est peut-être le plus intéressant est l'accès onomasiologique aux collocations : grâce aux FL il est possible de trier les collocatifs selon leur sens, car chaque FL exprime un rapport sémantique entre une base et son collocatif. Par exemple, la FL Real correspond à des collocatifs verbaux qui ont un sens de réalisation du « but » ou de la « fonction » du lexème qui constitue la base de la collocation. Donc, pour afficher tous les verbes qui expriment un sens de réalisation par rapport au lexème DIRITTO, on pourrait utiliser l'expression régulière suivante : *LexF = "Real.*_diritto.*"*

3.2 Difficultés

La difficulté principale rencontrée dans ce travail est l'absence de ressources lexicographiques de type Sens-Texte pour l'italien, ce qui a demandé un long travail de réflexion et de recherche contrastive italien-français pour associer les collocatifs aux FL ; en cas de doute sur l'association entre collocatif et FL, il n'y a pas de références pour nous guider.

Certains collocatifs verbaux de DIRITTO semblent correspondre à des FL différentes selon le contexte : le verbe BENEFICIARE en est un exemple. *Beneficiare di un diritto* ('bénéficier d'un droit') peut signifier, selon les contextes, 'avoir un droit' (ce qui correspond à la FL Oper1) ou bien 'exercer un droit', où on voit un sens de réalisation qui correspond à la FL Real1 (Mel'čuk, 1996). Les occurrences de ce collocatif dans le corpus semblent pointer vers ce deuxième cas, mais la distinction reste très fine et tracer des distinctions sémantiques nettes n'est pas toujours facile. Une annotation de ce type, même si elle se fonde sur la catégorisation formelle de Mel'čuk, qui se veut rigoureuse, pose des problèmes au moment d'interpréter les catégories sémantiques pour les appliquer à des cas concrets : le risque que deux annotateurs indépendants soient en désaccord sur l'annotation de certains lexèmes est très concret. C'est précisément pour cette raison que nous voyons la nécessité d'une étude sémantique complète et approfondie sur les FL et sur leur application dans le traitement des collocations de l'italien ; cette étude constituerait une référence précieuse pour un travail d'annotation sur l'ensemble des lexèmes d'un corpus.

Une autre difficulté importante a été l'annotation manuelle du corpus. L'automatisation du processus est envisageable (il serait possible de partiellement automatiser le processus en ayant recours au bootstrapping), mais elle ne sera possible que dès que des ressources lexicographiques Sens-Texte seront disponibles comme référence pour l'italien.

4 Conclusion

L'annotation présentée dans cette étude est une proposition méthodologique ; portant sur un seul lexème, les conclusions qu'on peut en tirer sont limitées, mais intéressantes. Par exemple, la présence d'information sémantique explicite dans l'annotation offre la possibilité d'accès onomasiologique aux collocations d'un corpus : grâce aux étiquettes sémantiques, il est possible d'affiner la recherche et de trier les collocatifs qui expriment un certain sens par rapport à un lexème donné. Ce type d'annotation apporterait d'autres bénéfices, surtout pour les étudiants :

- apprentissage plus systématique des collocations, grâce à la possibilité de les filtrer par sens, de les observer en contexte et d'observer leurs caractéristiques et leur patron syntaxique ;
- données textuelles réelles, non filtrées par la sélection du lexicographe ;
- ce type de ressource encourage l'étudiant à analyser les données de façon autonome et à réfléchir aux résultats, au lieu d'accepter passivement les informations fournies par des outils comme les dictionnaires ;
- avec le support des gloses, qui expliquent et vulgarisent la notation formelle, l'utilisateur pourrait se servir de l'annotation sans connaître en détail le système des FL.

Pour ces raisons, nous estimons que ce type d'annotation serait un outil efficace pour l'apprentissage des collocations (surtout pour les futurs traducteurs et interprètes), mais aussi pour la consultation.

Une objection possible est que l'emploi des corpus dans l'apprentissage linguistique est difficile pour les étudiants, et que l'annotation proposée ici risque de compliquer la tâche plutôt que de la faciliter. À cette objection, nous pourrions répondre que l'emploi des corpus en didactique est commun depuis au moins une décennie, et que de nombreuses études en démontrent l'efficacité (Bernardini *et al.*, 2003 ; Sinclair, 2004 ; Gallego-Hernandez et Rodriguez-Inés, 2016). Les corpus peuvent également constituer d'excellents outils pour l'apprentissage individuel, et leur utilisation autonome de la part des étudiants doit être encouragée, comme Bernardini (2004) le souligne.

De surcroît, en accord avec Mel'čuk (1998, 2003), Polguère (2000, 2003), L'Homme (2009, 2010) et Primož (2011), nous estimons que l'introduction des FL et de notions de linguistique Sens-Texte dans la didactique des langues apporterait d'importants bénéfices, en particulier dans la formation des futurs traducteurs et interprètes, car « appliquées de façon systématique et cohérente dans un champ lexical donné, le système des FL peut aider les étudiants en traduction à mieux maîtriser l'usage des collocations » (Primož, 2011 :129).

Bibliographie

Anctil, D. (2005), « Maîtrise du lexique chez les étudiants universitaires : typologie des problèmes lexicaux et analyse des stratégies de résolution de problèmes lexicaux », Mémoire de Maîtrise. Université de Montréal. <https://papyrus.bib.umontreal.ca/xmlui/handle/1866/16460> [consulté le 27.02.2018]

Augustyn, M., Tutin, A. (2009), « Constitution d'un corpus annoté autour du lexique des émotions : collocations et fonctions lexicales », in Beck D., Gerdes K., Miličević J. et Polguère A. *MTT' 09*, Jun 2009, Montreal, Canada. p. 25-35, 2009.

- Bernardini, S. (2004), « Corpora in the classroom. An overview and some reflections on future developments », in Sinclair J. (dir.) *How to Use Corpora in Language Teaching*, John Benjamins, Pays-Bas, ISBN-13 : 978-9027222824, p. 15-36. DOI : 10.1075/scl.12
- Bernardini, S., Stewart, D., Zanettin, F. (dir.) (2003), *Corpora in translator education*, Routledge, Royaume Uni, ISBN : 1900650606, 156 p.
- Carini, I. (2018), « Teoria Senso-Testo e Funzioni Lessicali. Una proposta per il trattamento delle collocazioni linguistiche », *Mémoire de Maîtrise*. Université di Bologna. [consulté le 21.03.2018]
- Castagnoli S., Lebani G.E., Lenci A., Masini F., Nissim M., Piunno V. (2015), « Towards a corpus-based online dictionary of Italian word combinations : the CombiNet project ». *COST ENeL WG3 meeting « Automatic Knowledge Acquisition for Lexicography »*, Herstmonceux Castle (UK), 13 August 2015.
- Evert, S. (2008). « Corpora and collocations », in Lüdelingand A., Kytö, M. (dir.), *Corpus Linguistics. An International Handbook*, 2, de Gruyter, Berlin, 1212-1248.
- Ferraresi, A., Bernardini, S. (sous presse), « Building EPTIC : A many-sided, multi-purpose corpus of EU Parliament proceedings » in Sánchez Nieto, M., Doval I. (dir.) *Parallel Corpora for Contrastive and Translation Studies*, John Benjamins, Pays-Bas, ISBN : 9789027202345, p. 123-139.
- Fontenelle, T. (1998), « Discovering significant lexical functions in dictionary entries », in Cowie A. P. (dir.), *Phraseology : theory, analysis and applications*, Oxford University Press, États-Unis, ISBN : 9780198299646, p. 189-207.
- Fontenelle, T., *Turning a Bilingual Dictionary into a Lexical-Semantic Database*. Niemeyer Verlag, coll. « Lexicographica/Series maior », Allemagne, 1997, ISBN : 3484309792, 348 p.
- Frassi, P., Tremblay, O. (2016), « Il Réseau Lexical du Français : une banque dati per l'apprendimento del lessico francese », in Bianchi, F., Leone, P., *Linguaggio e apprendimento linguistico. Metodi e strumenti tecnologici*. Officinaventuno, coll. « Studi AiTLA », Italie, ISBN : 8897657125 p. 155-172.
- Gallego-Hernandez, D., Rodriguez-Inés, P., dir. (2016), « Special Issue : Corpus Use and Learning to Translate, almost 20 years on », *Revue Cadernos de Tradução*, 36(1), ISSN 2175-7968, DOI : <http://dx.doi.org/10.5007/2175-7968.2016v36nesp1p9>
- Grossmann, F., Tutin, A. (dir.) (2003) *Les Collocations. Analyse et traitement*, De Werelt, coll. « Travaux et Recherches en Linguistique Appliquée » France, ISBN : 2-00-172701-1, 142 p.
- L'Homme, M.C. (2009), « A Methodology for Describing Collocations in a Specialized Dictionary », in Nielsen, S., Tarp, S. (dir.) *Lexicography in the 21st Century – In honour of Henning Bergenholtz*, John Benjamins, Pays-Bas, ISBN : 9789027223364, p. 237-256, DOI :
- L'Homme, M.C. (2010), « Designing Terminological Dictionaries for Learners based on Lexical Semantics : The representation of actants », in Fuertes-Olivera, P. (dir.) *Specialised Dictionaries for Learners*, De Gruyter, coll. « Lexicographica. Series Maior », Allemagne, ISBN : 978-3-11-023133-5, p. 141-153.
- Leech, G. (2005), « Adding Linguistic Annotation », in Wynne M. (dir.) *Developing Linguistic Corpora : a Guide to Good Practice*. Oxbrow Books, Royaume-Uni, ISBN : 9781842172056, p. 17-29.
- Lo Cascio, V., *Dizionario Combinatorio Italiano*, John Benjamins, Pays-Bas, 2013, ISBN 9789027212023, 1392 p.
- Lux-Pogodalla, V., Polguère, A. (2011), « Construction of a French Lexical Network : Methodological Issues », *First International Workshop on Lexical Resources, WoLeR 2011*, Aug 2011, Ljubljana, Slovenia. p. 54-61, 2011.
- Marengo, S., Robichaud, B. (2017), « Une nouvelle ressource pour la description des collocations : le DicoLiLex, dictionnaire de liens lexicaux », in *Représentations du sens linguistique VII (RSL VII)*. Sherbrooke, Canada.
- Mel'čuk, I. A. (1996), « Lexical functions : a tool for the description of lexical relations in a lexicon », in Wanner L. (dir.) *Lexical functions in Lexicography and natural language processing*, John Benjamins, Pays-Bas, ISBN : 9789027230348, p. 37-102, DOI :
- Mel'čuk, I. A. (1998), « Collocations and Lexical Functions », in Cowie P. (dir.) *Phraseology : theory, analysis and applications*, Oxford University Press, États-Unis, ISBN : 9780198299646, p. 23-53.
- Mel'čuk, I. A. (2003), « Les collocations : définition, rôle et utilité », in Grossmann F., Tutin A. (dir.) *Les Collocations. Analyse et traitement*, De Werelt, coll. « Travaux et Recherches en Linguistique Appliquée » France, ISBN : 2-00-172701-1, p. 23-31.
- Mel'čuk, I. A. (2013). *Semantics : from meaning to text*, vol. 1, John Benjamins, Pays-Bas, 2013, ISBN : 9789027205964, 436 p., DOI : <https://doi.org/10.1075/slcs.129>

Mel'čuk, I. A., Polguère, A., *Lexique actif du français : l'apprentissage du vocabulaire fondé sur 20 000 dérivations sémantiques et collocations du français*, De Boeck, Belgique, 2007, ISBN : 978-2801113455, 525 p.

Mel'čuk, I., Arbatchewsky-Jumarie, N., Iordanskaja, L., Mantha, S., Polguère, A., *Dictionnaire explicatif et combinatoire du français contemporain : Recherches lexico-sémantiques I, II, III, IV*, Presses de l'Université de Montréal, Canada, 1984, 1988, 1992, 1999, ISBN : 9782760617384

Nesselhauf, N., *Collocations in a learner corpus*, John Benjamins, Pays-Bas, 2015, ISBN : 9789027222855, 332 p., DOI : <https://doi.org/10.1075/scl.14>

Polguère, A. (2000), « Une base de données lexicales du français et ses applications possibles en didactique », *Revue de linguistique et de didactique des langues (Lidil)*, N. 21, p. 75-97

Polguère, A. (2003), « Collocations et fonctions lexicales : pour un modèle d'apprentissage », in Grossmann F., Tutin A. (dir.) *Les Collocations. Analyse et traitement*, De Werelt, coll. « Travaux et Recherches en Linguistique Appliquée » France, ISBN : 2-00-172701-1, p. 117-133.

Primož, J. (2011), « Meaning-Text Theory in the translator's classroom », *International Journal of Translation*, N.13, ISSN : 1722-5906, p. 129-138

Smadja, F. (1991), « Macro-coding the lexicon with co-occurrence knowledge » in Zernik (dir.) *Lexical Acquisition : Exploiting On-Line Resources to Build a Lexicon*. Hillsdale, États Unis, ISBN : 9780805811278, p. 165-189.

Tiberii, P., *Dizionario delle collocazioni : le combinazioni delle parole in italiano*, Zanichelli, Italie, 2012, ISBN : 978-8808220936, 640 p.

Tremblay, O. (2009), « Une ontologie des savoirs lexicologiques pour l'élaboration d'un module de cours en didactique du lexique », Thèse de doctorat, Université de Montréal, <http://olst.ling.umontreal.ca/pdf/OphelieTremblayThese2009.pdf> [consulté le 23.01.2018]

Tremblay, O., Polguère, A. (2014), « Une ontologie linguistique au service de la didactique du lexique », in Neveu, F., Blumenthal, P., Hriba, L., Gerstenberg, G., Jmeinschaefer, J., Prévost, S., 4^e *Congrès Mondial de Linguistique Française (CMLF 2014)*, Jul 2014, Berlin, Allemagne, EDP Sciences, p. 1173-1188, 2014, DOI : 10.1051/shsconf/20140801383

Tutin, A. (2005), « Annotating Lexical Functions in Corpora : Showing Collocations in Context », *2nd International Conference on the Meaning-Text Theory*. Moscow, June 23-25 2005, <http://courses.cs.tau.ac.il/0368-4341/shared/Papers/Collocations/MTT.pdf> [consulté le 23.01.2017]

Urzì, F., *Dizionario delle combinazioni lessicali*. Convivium, Luxembourg, 2009, ISBN : 978-2-9599889-7-4, 998 p.