Terminologie, Intelligence Artificielle et Psychologie cognitive

Anne Condamines

Terminologie, intelligence artificielle, psychologie cognitive : réflexions sur les interactions possibles dans l'étude de la variation en langue spécialisée

Anne Condamines

Résumé : Cet article présente une réflexion sur les possibilités de prise en compte, par les approches en IA, de la variation en terminologie liée à une émotion, en prenant appui sur les relations existant entre IA, terminologie et psychologie cognitive. Après le descriptif des relations entre ces trois disciplines du point de vue de la prise en compte de la situation extralinguistique, deux études sont présentées, l'une sur la présence *vs* absence de préposition dans la construction de *pêcher* et *rivière* (en français) et *to fish* et *river* (en anglais), l'autre sur l'environnement lexical de chaque structure en anglais. Les résultats montrent que la structure sans préposition est utilisée principalement dans les sites de pêcheurs contenant l'expression d'une subjectivité (forums, blogs…) et que l'environnement lexical relève de catégories sémantiques différentes pour les deux structures (dont l'expression du bien-être dans le corpus « sans préposition »). Les résultats de ces deux études mettent en évidence des différences de choix de construction syntaxique de la part des pêcheurs qui semblent corrélées avec un rapport particulier, affectif, avec la rivière. L'article se termine par une évaluation concernant les apports et les limites de l'IA dans des études visant à prendre en compte des éléments extralinguistiques dans la description des phénomènes de variation dans des corpus spécialisés.
Mots-clés : émotion, intelligence artificielle, psychologie cognitive, terminologie, variation

Abstract: This article presents a reflection on the possibilities of taking into account the variation in terminology linked to an emotion by AI approaches; it is based on the existing relations between AI, terminology and Cognitive Psychology. After the description of the relations between these three disciplines, taking into account the extra-linguistic situation, two studies are presented, one on the presence *vs*. absence of preposition in the construction of *to fish* and *river* (in English) and *pêcher* and *rivière* (in French), the other on the lexical environment of each structure in English. The results show that the structure without the preposition is mainly used in anglers' sites displaying an expression of subjectivity (forums, blogs...) and that the lexical environment belongs to different semantic categories (including the expression of well-being in the "without preposition" corpus). The results of these two studies highlight differences in the choice of syntactic construction on the part of anglers, which seem to correlate with a particular emotional relationship with the river. The article concludes with an evaluation of the contributions and limitations of AI in studies aiming at taking into account extra-linguistic elements in the description of variation phenomena in specialized corpora.
Keywords : emotion, artificial intelligence, cognitive psychology, terminology, variation

Introduction

Après une période de rapprochement fructueuse en lien avec l'analyse automatique de corpus spécialisés, au début des années 1990, les relations entre l'intelligence artificielle (IA) et la terminologie se sont beaucoup distendues lors des dernières années, avec le développement de la sémantique distributionnelle et de l'apprentissage profond dans le TAL (Traitement Automatique de la Langue). En IA, la construction et la mise à jour des ontologies, apparentées à des réseaux terminologiques, se font désormais, le plus souvent, sans que soit fait appel à des connaissances linguistiques. Dans le même temps, certains psychologues ont cru voir, dans les (parfois très bons) résultats de l'IA obtenus grâce à des méthodes dites de sémantique distributionnelle, une confirmation de leurs hypothèses sur

l'apprentissage et la mise en œuvre du lexique mental, qui se feraient indépendamment de situations particulières. Cette position pose question à la fois pour la recherche en terminologie et pour la recherche en psychologie cognitive. D'une part, les études en terminologie, qui, par essence, prennent en compte une situation liée à une connaissance spécialisée, s'intéressent de plus en plus à la notion de variation, en fonction de différents paramètres extralinguistiques. D'autre part, de nombreux chercheurs en psychologie cognitive plaident en faveur d'une cognition incarnée et située. Les relations entre ces trois disciplines sont donc complexes. Tout en tenant compte de cette complexité, cet article s'interroge sur les possibilités de complémentarité entre ces disciplines, avec un point de vue orienté vers les besoins en analyse de la variation dans les corpus spécialisés.
Il est d'abord fait un état des lieux des relations bilatérales entre les trois disciplines : terminologie et intelligence artificielle, intelligence artificielle et psychologie cognitive, psychologie cognitive et terminologie (1er paragraphe). Dans le paragraphe 2, l'étude d'un phénomène particulier, l'alternance de construction (directe *vs via* une préposition) de *pêcher* et de *rivière*, montre comment la variation peut intervenir dans les discours spécialisés en fonction de l'implication affective du pêcheur, et s'interroge sur les possibilités que ce type de variation soit repéré par des méthodes d'IA.

1. Des relations bilatérales

Cette partie rend compte des relations existant entre les trois disciplines : terminologie, IA et psychologie cognitive afin de poser le cadre de la réflexion. Dans le dernier sous-paragraphe 1.4, nous mettons en question la façon dont l'IA emprunte la notion de « sémantique distributionnelle » à la linguistique pour justifier son approche.

1.1 Terminologie et IA

À la fin des années 1980, la terminologie et l'ingénierie des connaissances (IC) (un des aspects de l'intelligence artificielle) se sont rapprochées sur la base de différents constats.

a) Les deux disciplines utilisaient des modes de représentation de la connaissance constitués de réseaux de concepts reliés par des relations. Termes et relations étaient des mots (voire des groupes de mots le plus souvent).
b) Les deux disciplines se sont orientées vers l'utilisation de corpus pour repérer les termes et les relations. Cela permettait à l'IC de ne pas construire les systèmes à base de connaissances à partir des seuls entretiens avec des experts (pas toujours à l'aise dans cet exercice). Quant à la terminologie, elle pouvait bénéficier des développements de la linguistique de corpus et utiliser des méthodes similaires à celles utilisées pour la lexicographie à base de corpus.
c) Dans les deux cas, les méthodes de construction des réseaux de concepts se basaient sur la mise en œuvre de connaissances linguistiques, en particulier des marqueurs de relations (en lien avec la notion de « *knowledge rich context* » proposée par Meyer (2001). Pour mémoire, les marqueurs de relations sont des éléments lexico-syntaxiques qui permettent de repérer,

plus ou moins systématiquement, des relations en corpus. Par exemple : [Tous les N1 sauf les N2] permet de repérer une relation d'hyperonymie entre N1 et N2.

1) Tous les poissons sauf les truites sont relâchés.

Ce rapprochement de l'IA et de la terminologie s'est manifesté en particulier par la création des « bases de connaissances terminologiques » (Meyer *et alii* 1992 ; Condamines 2018c) et par le développement des études sur les marqueurs de relations conceptuelles et de leur variation en contexte, par exemple en fonction du genre textuel (Auger, Barrière 2008 ; Condamines 2002 ; Marshman *et alii* 2008).
En France, le groupe TIA (Terminologie et Intelligence Artificielle) créé par Didier Bourigault et Anne Condamines en 1993 a permis à des chercheurs de différentes communautés scientifiques : linguistique, terminologie, informatique, sciences de l'information, de se réunir et d'interroger les rapprochements possibles autour du thème général « corpus et terminologie » (Aussenac, Condamines 2007).
Depuis une dizaine d'années, les deux disciplines se sont éloignées. La terminologie, avec le développement de la terminologie textuelle, a évolué vers la prise en compte d'autres objectifs que la seule construction de réseaux de concepts à partir de textes (Condamines 2018 ; Condamines, Picton à paraître). Quant à l'IA, c'est moins la construction de réseaux conceptuels (ontologies) en tant que tels qui devient l'objectif des études, mais plutôt des applications finales, utilisant (ou pas) des ontologies : recherche d'information, traduction, question-réponse, recherche de thématiques… (Turney, Pantel 2010). L'intérêt commun entre les deux communautés (IA et terminologie), en particulier pour la recherche et l'étude des relations conceptuelles, s'est donc largement émoussé et les disciplines se sont éloignées.

1.2 Psychologie cognitive et IA

Le développement des méthodes d'apprentissage du sens mettant en œuvre la « sémantique distributionnelle » *via* la construction de vecteurs en IA (il faut le reconnaître, avec un succès certain du point de vue des applications) a ravivé les discussions entre différents courants de la psychologie cognitive à propos du « lexique mental » (Segui 2015). D'une part, il a montré l'importance du rôle de l'environnement syntaxique des mots (leur distribution) pour l'élaboration du sens, ce qui n'était pas forcément un aspect reconnu en neuropsychologie. Mais, d'autre part, il a pu laisser penser que l'apprentissage des mots se faisait par la mémorisation de ces seuls contextes langagiers, syntaxiques (collocations), indépendamment des situations dans lesquelles ils apparaissent.

> *Cognitive scientists have argued that there are empirical and theoretical reasons for believing that VSMs [Vector Space Model of Semantics], such as LSA and HAL, are plausible models of some aspects of human cognition* (Turney, Pantel 2010 : 144).

Cette hypothèse est fortement contestée par les tenants de la cognition incarnée (*grounded cognition*) (Barsalou 2003), qui estiment que l'apprentissage des mots ne peut se faire sans la prise en compte de la situation extralinguistique mais aussi des réactions sensori-motrices des locuteurs/ auditeurs.

> *We draw two conclusions […]. The first is that high-dimensional theories such as LSA and HAL are inadequate accounts of human meaning because the symbols (high dimensional vectors) are not grounded* (Glenberg, Robertson 2000 : 384).

Les méthodes de traitement automatique de la langue inspirées de la sémantique distributionnelle tiennent le plus souvent à l'écart cette « inscription corporelle » de la connaissance, certains auteurs remettant carrément en cause cette dimension.

> *The importance of embodiment and grounding is exaggerated, and the implication that there is no highly abstract representation at all, and that human-like knowledge cannot be learned or represented without human bodies, is very doubtful* (Landauer 1999 : 624).

Cet éloignement est expliqué de la façon suivante par Glenberg et Robertson : « *The reason for using ungrounded symbols is clear : they are far easier to use in computer and mathematical simulations than are grounded representations* » (Glenberg, Robertson 2000 : 399).Toutefois, des études tentent de montrer que les deux approches concernant la connaissance lexicale ne sont pas incompatibles dans la perspective de la sémantique distributionnelle au sein de l'IA ; la prise en compte de la dimension incarnée est ainsi en train de se développer :

> *There is a growing trend in cognitive sciences to find a common ground in which embodied cognition and distributional approaches to meaning could eventually meet* (Lenci 2008 : 25).

1.3 Terminologie et psychologie cognitive

Nous l'avons vu, la réflexion sur les relations entre terminologie et intelligence artificielle s'est développée à partir des années 1990. Pour une part, cet état de fait, qui a encouragé la terminologie à travailler sur des données réelles (des corpus) est lié à une réaction contre la vision prescriptive de la terminologie proposée par Wüster dans la Théorie Générale de la Terminologie à partir des années 1930 (*General Theory of Terminology*, GTT) (Wüster 1985). Avec la GTT il s'agissait d'établir des référentiels terminologiques, sorte de normes permettant une diffusion espérée transparente des informations, entre entreprises d'un domaine, dans une langue ou entre langues. Aussi louable qu'il soit, ce point de vue est peu compatible avec le fonctionnement discursif, qui, par essence, se libère parfois des normes, en fonction du contexte (partage de la connaissance dans des communautés restreintes), en fonction du besoin (efficacité de la communication), de l'évolution dans le temps, de la nécessité de créer des variantes en fonction de l'apparition de nouveaux concepts, etc. Plusieurs auteurs se sont opposés à cette vision prescriptive. Certains se sont inscrits dans la perspective de la linguistique de corpus outillée (la terminologie textuelle par exemple). D'autres dans la perspective de la théorie des *frames* (Faber 2012). D'autres chercheurs se sont appuyés sur la sociologie comme la socioterminologie (Gaudin 2003), d'autres encore à la fois sur la sociologie et la psychologie comme la terminologie sociocognitive (Temmerman 2000). Cette prise en compte des aspects sociologiques et psychologiques du fonctionnement terminologique a marqué le rapprochement avec les travaux en linguistique sociocognitive (Kristiansen, Dirven 2008) et en linguistique de corpus (Gries 2015). La variation des termes, entendus comme des unités (ou des poly-unités) lexicales, ne pouvait donc plus être ignorée.
Quant à la dimension « affective » de la terminologie, elle n'est que très rarement prise en compte, ce que regrettent des auteurs comme Baumann :

> […] *theories of emotion which have been ignored by LSP research for a long time are of increasing methodological and methodical significance because they offer far-reaching strategic orientations for the communicative-cognitive analysis of information processing in LSP texts*. (Baumann 2007 : 322).

Cette dimension émotionnelle fait écho à la dimension incarnée de la psychologie cognitive.

1.4 Linguistique et IA : la question de la « sémantique distributionnelle »

La plupart des travaux en TAL se revendiquent actuellement d'une approche distributionnelle. Or, dans sa prise en compte par les outils, ce point de vue est en partie amputé des propositions des origines.
Depuis les premières réflexions sur le distributionnalisme dans les années 1930, inspirées par Bloomfield, l'approche distributionnelle en linguistique a beaucoup évolué. À partir des années 1950, deux principaux courants sont apparus, l'un, américain, avec comme chef de file Harris (Harris 1954) et l'autre, anglais, avec comme chef de file Firth (Firth 1957). La méthode harissienne, inspirée par une vision behavioriste du fonctionnement du sens, préconise une approche mathématique pour arriver, par différents types d'opérations, au sens intrinsèque d'une phrase. L'approche anglaise est beaucoup plus ancrée dans une vision sociologique (voir ci-dessous). C'est d'abord l'approche harissienne qui a été convoquée en TAL pour justifier les travaux sur le sens en lien avec l'accès à de très gros volumes de données textuelles. Mais, tout l'appareillage mathématique de la méthode harissienne n'étant pas mis en œuvre dans cette approche en TAL (on a d'ailleurs plutôt parlé d'une méthode « à la Harris »), c'est l'autre courant de l'analyse distributionnelle qui a ensuite été convoqué, celui de Firth. On ne compte ainsi plus le nombre de travaux de TAL en sémantique distributionnelle qui se revendiquent de Firth en citant cet extrait : « *You shall know a word by the company it keeps* » (Firth 1957 : 11). Le problème est que cette phrase est isolée du cadre général de l'approche firthienne. Un autre extrait de Firth est, à cet égard, parlant :

> *First the structure of the appropriate contexts of situation must be stated. Then the syntactical structure of the texts. The criteria of distribution and collocation should then be applied* (Firth 1968 [1952] : 19).

Ainsi, pour Firth (et l'école londonienne), la situation de communication contribue à la construction du sens et joue même un rôle déterminant. Notons d'ailleurs que la situation est aussi prise en compte dans la proposition de Harris, qui a créé le concept de sous-langage, sous-partie de la langue associée à des domaines de connaissances particuliers. Même si les études évoluent, en particulier dans l'analyse de corpus spécialisés (Fabre *et alii* 2014), la situation de communication est assez peu prise en compte dans les travaux en TAL qui s'appuient sur la sémantique distributionnelle. La principale explication est liée au fait que les méthodes d'apprentissage profond se mettent en place sous la forme de constitution de vecteurs (Heylen, Bertels 2016), ce qui nécessite des volumes de données très importants et que de telles quantités de données ne sont pas toujours disponibles pour les domaines spécialisés (domaines parfois très restreints ou bien dont les documents sont confidentiels) (Boleda 2020).
Pour cette raison, et d'autres, les outils de TAL basés sur l'apprentissage profond peuvent avoir des difficultés à prendre en compte la variation dans les domaines spécialisés.

2. Étude de cas

Cette étude va nous permettre de rendre compte d'un phénomène de variation concernant un fonctionnement lexico-syntaxique dans un domaine spécialisé, celui de la pêche. Nous verrons ensuite dans quelle mesure une étude de ce type pourrait être assistée par des méthodes d'IA.

2.1 La problématique

Nous abordons ici la présentation d'une étude dans laquelle internet a été utilisé comme corpus, dans une approche en partie outillée.
Il s'agit de l'étude de la variation de la construction entre *pêcher* et *rivière(s)*, avec ou sans préposition. C'est par hasard que nous avons rencontré la construction sans préposition (*j'ai déjà pêché cette rivière*). Cet énoncé nous a paru échapper à la norme qui voudrait que *pêcher* soit suivi d'une préposition (*en, dans, sur*) pour introduire le complément de lieu. D'emblée, ce choix de construction nous a semblé propre aux pêcheurs et, par ailleurs, marquer une volonté inconsciente de rendre compte d'un lien privilégié avec la rivière, ce qui justifiait la suppression de la préposition. Cette hypothèse d'une variation sémantique en lien avec une variation de forme fait écho à celle de la grammaire constructionnelle : « *[…] in construction-based theories, grammar consists of a structured inventory of pairings of form and meaning* » (Goldberg 1996 : 8).

2.2 Deux types d'analyses

Deux types d'analyse ont été menés en utilisant les données disponibles sur Internet (l'exploration a débuté au cours de l'année 2012).
Dans la première étude, nous avons recherché toutes les occurrences correspondant à [pêcher + préposition + (déterminant) + rivière(s)] ou [pêcher + déterminant + rivière(s)]. Seules les formes à l'infinitif ont été recherchées pour limiter le nombre d'occurrences à traiter. Les déterminants pouvaient être un défini ou un indéfini : *le, la, les, un, une, des* et les prépositions *dans, en, sur*.
À chaque occurrence, nous avons associé un type de site internet : « pêche et subjectivité » ou « autre ».
Cette analyse a aussi été réalisée pour l'anglais en recherchant les structures [*to fish* + (déterminant) *river(s)*] et [*to fish* + préposition + (déterminant) + *river(s)*] avec préposition = *on*, *within*, *ou in* et déterminant = *a* ou *the*.
Dans la seconde analyse, nous nous sommes focalisée sur l'anglais et nous avons constitué deux sous-corpus, l'un avec les phrases contenant la structure avec préposition, l'autre avec les phrases contenant la structure sans préposition. Pour mettre en œuvre une approche lexicométrique, nous avons utilisé le logiciel AntConc (Anthony 2014) pour comparer les *keywords* significativement plus présents dans chacun des sous-corpus. Nous avons ensuite catégorisé sémantiquement « à la main » l'ensemble de ces *keywords* afin d'essayer d'établir un environnement sémantique caractéristique de chacune des constructions (avec *vs* sans préposition).

2.3 Résultats

Le tableau 1 rend compte de la répartition des structures avec préposition *vs* sans préposition, en français et en anglais, en fonction de la nature des sites.

	Tous les sites	Sites de pêche avec subjectivité

	Français	Anglais	Français	Anglais
Structures avec préposition	82,3 %	49,7 %	61,6 %	30,8 %
Structures sans préposition	17,7 %	50,3 %	38,2 %	69,2 %

Tableau 1 : Structures avec *vs* sans préposition dans tous les sites *vs* dans les seuls sites de pêche avec subjectivité, pour le français et l'anglais

Les résultats présentés dans ce tableau confortent notre hypothèse de départ.
Pour le français, dans le corpus global, les occurrences sans préposition ne sont pas rares (17, 7 %), contrairement à ce qu'un locuteur du français standard (c'est-à-dire non pêcheur) pourrait penser. On peut constater de surcroît que la nature du site joue un rôle dans l'apparition d'une ou de l'autre structure ; en effet, dans les sites de pêche avec une dimension subjective, la construction directe est utilisée dans 38,2 % des cas. Précisons qu'un même site peut utiliser les deux structures. La mise en œuvre du Chi2[1] a confirmé l'existence d'une corrélation entre nature du site et apparition de la structure avec *vs* sans préposition.
Pour l'anglais, on peut noter que la structure sans préposition est présente presque à égalité avec la structure avec préposition (50,3 % *vs* 49,7 %). Les deux structures sont d'ailleurs admises et enregistrées dans les dictionnaires anglais. Par ailleurs, tout comme pour le français, la nature du site est en lien avec le choix d'une ou de l'autre structure : les sites de pêche avec une dimension subjective favorisent la présence de la construction sans préposition.

2) Voici quelques exemples de phrases avec ces deux constructions, pour le français et pour l'anglais.Très jolie rivière ! tu as du bol de pouvoir pêcher une rivière aussi sauvage.
3) Avant de pêcher dans la rivière Northwest, les pêcheurs à la ligne doivent se procurer un permis de pêche du saumon du parc national.
4) I love to fish rivers, every single one I have ever been on is different.
5) To fish in the River Fowey you will need an Environment Agency Licence, (except under 12's), these can be purchased at a Post Office or by Telephone.

Pour ce qui concerne la seconde étude (menée seulement sur l'anglais), les résultats sont aussi très intéressants. Le tableau 2 rend compte des résultats chiffrés obtenus avec AntConc (Anthony 2014) pour les deux sous-corpus anglais. Rappelons que l'un de ces sous-corpus est constitué des paragraphes contenant le verbe « pêcher » suivi d'une préposition, l'autre des paragraphes contenant le verbe « pêcher » construit directement avec *rivière(s)*.
Le nombre de mots, mais aussi le nombre de lemmes, sont proches dans les deux corpus, ce qui a facilité la comparaison statistique.

	Corpus sans préposition	Corpus avec préposition
Nombre de mots	41361	40365

[1] Le Chi2 est un test statistique permettant de mesurer l'existence d'une dépendance entre deux variables.

Nombre de lemmes	2715	2440
Nombre de lemmes significatifs (avec *keyness* > 3.84)	322	319

Tableau 2 : Résultats quantitatifs pour l'étude comparée des deux sous-corpus anglais

Parmi les lemmes spécifiques à l'un ou l'autre corpus, nous avons pu identifier 10 catégories sémantiques, 5 pour chacun des corpus.
Voici des exemples de lemmes associés à chaque catégorie ; pour le corpus « sans préposition » :

- Mois ou saisons : *January, February, spring, summer,*
- Poissons : *trout, pike, grayling, walleye, bream*,
- États ou régions : *Normandy, Nevada, Alabama,*
- Vocabulaire positif : *inspiring, ideal, beautiful, clarity, peacefulness*,
- Accessoires de pêche : *tackle, accessories, wader, bait, nymphs, line, braids* ;

pour le corpus « avec préposition »

- Vocabulaire légal : *permission, unlawful, license, permit, law*,
- Vocabulaire économique : *property, owner, landowner, leaseholder*,
- Vocabulaire relevant du danger : *chemical, lethal, polluted, danger, arsenic, decrease, threat*,
- Éléments naturels : *cormorant, flower, animal, reef, plant, crocodile*,
- Relations familiales : *grandchildren, husband, ancestors, family*.

Les extraits 6) et 7) sont deux nouveaux exemples illustrant les deux catégories d'usage :

6) Should you be lucky enough to have the opportunity to fish rivers or streams the best flies to start with would be wet flies and nymphs.
7) You need a permit to fish in the river, but I am unsure of the cost.

Certaines catégories peuvent paraître étonnantes. Ainsi la surreprésentation des éléments naturels et la surreprésentation des relations familiales dans le corpus « avec préposition » donc « général ». Ces résultats peuvent peut-être s'expliquer par le fait que les pêcheurs sont surtout focalisés sur la rivière comme élément naturel et qu'ils préfèrent pêcher seuls (ou, peut-être, avec des amis). En revanche, ils sont particulièrement sensibles à leur environnement situationnel : le lieu et le moment où ils pêchent, mais aussi leur équipement et les types de poissons qu'ils peuvent rencontrer.
Ces deux études, rapidement présentées (pour plus de détails, voir Condamines 2017, 2018a, 2018b, 2021) mettent en évidence deux aspects.
D'une part, la construction sans préposition est plus utilisée par les pêcheurs que par les non-pêcheurs. D'autre part, lorsque cette construction est choisie, elle s'accompagne d'un vocabulaire faisant intervenir la dimension affective ou émotionnelle. En témoigne en particulier l'utilisation d'un vocabulaire relevant du bien-être. Allant dans le sens de cette implication émotionnelle, on peut noter des énoncés dans lesquels la rivière est personnalisée, au point de faire l'objet de sentiments.

1) On a whim, I decided to fish a river in Oregon I had visited before, and fell in love with.

D'un point de vue méthodologique, notons que la recherche du lexique le plus significatif dans l'environnement des deux types de structures relève bien d'une analyse distributionnelle, focalisée sur le lexique présent dans l'entourage de chacune des structures, mais elle ne tient pas compte du rôle syntaxique de ces éléments.
Au-delà de la description linguistique, la description de ce phénomène a une importance pour comprendre le fonctionnement cognitivo-syntaxique des experts, qui n'est pas le fruit du hasard. Il faut donc aussi voir si ce phénomène de « transitivation du complément de lieu » se retrouve pour d'autres verbes et pour d'autres langues.

2.4 D'autres verbes, d'autres langues
2.4.1 D'autres verbes

En cherchant sur le web un peu au hasard, nous avons rencontré d'autres verbes admettant (même si c'est parfois très rare) la construction directe du complément de lieu, par exemple :

- Pagayer la rivière

8) Tout sourire et solidement complices, Manon et moi allions enfin pagayer la rivière Kanasuta.

- Plonger la rivière

9) Le chevalier Percevan va plonger la rivière souterraine qui alimente le puits où il était esclave.

- Skier la montagne

10) Profiter de la poudre pour skier la montagne de Lure est devenu un évènement rare.

- Chasser la lande

11) Henri VIII, qui aimait chasser la lande entourant le village, a donné le manoir à Anne Boleyn pour la vie.

Le cas de « *chasser* + nom de lieu » est particulièrement étudié dans Condamines (2018b).
En revanche, nous n'avons trouvé aucune occurrence de *tuer la rivière,* ni de *harponner la rivière* (avec des verbes qui peuvent se trouver à la place de *pêcher* devant *poisson*) ni de *danser une salle*, *boxer un ring*, *chanter La Scala/ L'Olympia*. Cette différence entre les verbes pourrait être liée à la nature du lieu : *montagne, rivière, lande* renvoient à des lieux ouverts contrairement à *ring, salle, Olympia*. Mais il faudrait des études plus poussées pour vérifier s'il s'agit d'un aspect décisif dans l'impossibilité de la construction.
Il se pourrait aussi que l'origine territoriale du locuteur joue un rôle. Pour le français, les Québécois pourraient, par exemple, utiliser plus facilement la construction sans préposition (l'exemple 9 est ainsi issu d'un site québécois). Plus généralement, Callies (2018) a noté, en anglais américain, une tendance à supprimer les prépositions par exemple dans *he graduated Stanford* au lieu de *he graduated from Stanford*. Il faudrait envisager une catégorisation de l'origine des sites internet pour pouvoir prendre en compte cette caractéristique, ce qui est loin d'être toujours faisable.

2.4.2 D'autres langues

Nous avons mené le même type d'étude, concernant *pêcher* et ses traductions, avec d'autres langues : l'espagnol, l'italien et l'occitan (en collaboration avec des locuteurs de ces langues). Le tableau 3 rend compte du nombre d'occurrences trouvées sur le web pour chaque structure dans différentes langues.

	Avec préposition	Sans préposition	Total
Anglais	1094	1108	2202
Espagnol	1226	384	1610
Français	1213	261	1474
Italien	961	26	987
Occitan	79	0	79

Tableau 3 : Présence *vs* absence de préposition dans les constructions dans différentes langues

12) Me cuesta abandonar sus orillas y mas el último día de temporada, sabiendo que no volveré a pescar el río hasta dentro de unos cuantos meses (espagnol).
13) I periodi migliori per pescare il fiume Snake sono all'inizio della primavera prima del ballottaggio (italien).
14) L'Ib-Salut esperarà esdeveniments i mai no jugarà la tàctica de pescar dins un riu remogut (occitan).

On peut noter que les deux constructions sont présentes pour toutes les langues hormis l'occitan. Pour cette langue, le nombre d'occurrences, même pour la construction canonique, est très faible (79). La question des langues régionales, faiblement présentes sur l'Internet, constitue un des problèmes majeurs des études linguistiques qui voudraient mettre en œuvre des méthodes d'apprentissage profond qui nécessitent des données volumineuses, nous l'avons déjà vu.
Dans une autre étude (Condamines 2017) nous avons pu voir que la nature du site (+ *pêche* + *subjectif*) jouait un rôle en espagnol comme en français et en anglais. En revanche, pour l'italien (langue pour laquelle la construction directe est rare), on ne retrouve pas cette corrélation.

2.5 L'IA pourrait-elle assister le travail des terminologues textuel(le)s ?

Dans cette dernière sous-partie, en prenant l'exemple de l'alternance de construction des compléments de lieu construits avec *vs* sans préposition, nous nous demandons quel pourrait être l'apport des systèmes d'IA basés sur des corpus volumineux. Du côté des apports possibles de l'IA, nous pouvons noter :

- La possibilité de rechercher les formes verbales non seulement à l'infinitif mais à tous les temps possibles, ce qui augmenterait le nombre de données utilisables. Il faudrait cependant pouvoir travailler sur un corpus étiqueté en parties du discours.
- La possibilité de faire porter la recherche sur d'autres verbes que le verbe « pêcher ».

- La possibilité de repérer automatiquement, pour un même verbe, les arguments construits directement *vs* construits indirectement.

La possibilité d'établir une corrélation entre nature de la construction et nature du site, avec un bémol concernant le fait qu'il peut être difficile d'établir précisément la nature du site (*i.e.* le genre textuel dont il relève). Ces deux derniers points, s'ils s'avéraient réalisables, constitueraient une réelle aide pour l'analyse linguistique.
Du côté des difficultés, nous pouvons noter :

- La nécessité de disposer d'un grand nombre de données ; or, nous l'avons vu pour l'occitan par exemple, certaines langues sont peu dotées du point de vue de la quantité de textes sur Internet.
- La difficulté pour les systèmes d'IA de repérer la nature sémantique des compléments qui sont construits avec ou sans préposition (par exemple le lieu dans le cas de *pêcher*). Des méthodes de sémantique distributionnelle en corpus général pourraient sans doute être convoquées pour contribuer à cette tâche.
- La nécessité, pour les linguistes-terminologues, de connaître les (très complexes car basés sur des savoirs mathématiques) systèmes d'IA et leur mode de programmation pour les adapter à une question nouvelle, ou bien de faire appel à des informaticien(ne)s spécialisé(e)s dans le TAL (ce qui les rend dépendant(e)s d'un tiers).
- Le fait que la question de la corrélation entre différents types de phénomènes (ici, la nature d'une construction et la nature du site) relève d'une intuition linguistique et pas d'une proposition de l'outil. De fait, le nombre des corrélations potentiellement repérées par des outils est très élevé et beaucoup n'ont pas de sens pour le (la) linguiste-terminologue. Imaginons par exemple que l'outil découvre que les pêcheurs utilisent plus la lettre « e » que les non pêcheurs. À cette corrélation, le (la) linguiste-terminologue ne peut donner aucune interprétation pertinente pour son point de vue. Et il n'y a sans doute aucune interprétation à donner.

En fait, on pourrait espérer qu'un outil contribue à répondre à la question : quels sont les verbes qui ont des constructions différentes (présence *vs* absence d'une préposition) pour les mêmes arguments et dans quelle mesure la nature du site joue-t-elle un rôle significatif dans cette alternative ? Cette question de la nature du site permettrait de prendre en compte des éléments extralinguistiques *via* la notion de genre textuel, voire des éléments relevant d'une cognition incarnée *via* l'étude du lexique. En outre, l'objet d'observation ainsi constitué se rapprocherait d'un corpus et donc d'un véritable objet d'étude linguistique et pas seulement de données car, comme le rappelle Rastier : « les données n'ont pas de sens » (Rastier 2021, 232). Mais les outils ne pourraient sans doute pas proposer tout seuls d'aller regarder du côté de telle corrélation, qui semble présenter un intérêt linguistique. En d'autres termes, l'IA ne peut pas (en tout cas pour l'instant) remplacer l'intuition des linguistes, c'est-à-dire des expert(e)s de la langue qui élaborent des hypothèses.

Conclusion

Cet article a eu pour but de dresser un panorama des relations entre IA, terminologie et psychologie cognitive afin de voir ce qu'elles peuvent apporter à la réflexion sur la prise en compte de la variation dans les langues spécialisées. La présentation de l'étude sur la variation de construction (avec préposition *vs* sans préposition) entre *pêcher* et *rivière* et *to fish* et *river* a montré un lien entre la nature des sites internet

et la nature de la construction. L'article a aussi montré que la dimension émotionnelle des experts (les pêcheurs) pouvait être repérée par la significativité sémantique de l'environnement lexical des deux structures (présence importante du lexique du bien-être dans les phrases contenant les structures sans préposition). Cet aspect va dans le sens des propositions de la psychologie cognitive concernant la cognition incarnée, c'est-à-dire la prise en compte de la situation dans l'apprentissage et le fonctionnement du lexique mental. Nous avons aussi essayé de montrer comment les nouvelles méthodes d'IA pouvaient (ou ne pouvaient pas) contribuer à la mise au jour de ces fonctionnements.
L'évolution rapide des méthodes d'IA utilisées en TAL, désormais surtout basées sur la sémantique distributionnelle et l'apprentissage profond, a abouti à des résultats spectaculaires, par exemple dans la traduction automatique. Mais elle a aussi déstabilisé les expert(e)s de la langue, qui se sont senti(e)s dépossédé(e)s de leurs connaissances et de leurs compétences puisque les outils fonctionnent sans connaissance linguistique. Passée cette étape de déstabilisation, les linguistes doivent réfléchir à la position qu'ils/ elles peuvent avoir par rapport à cet état de fait. Au-delà de la seule production de ressources pour alimenter les systèmes de TAL, clairement en perte de vitesse, les linguistes pourraient prendre en compte les possibilités de ce nouveau TAL afin d'évaluer s'ils/ si elles peuvent, et à quelles conditions, intégrer ces méthodes dans leur objectif qui est, *in fine*, de comprendre les fonctionnements langagiers (essentiellement sémantiques) pour répondre, éventuellement, à des besoins sociétaux.

Bibliographie

Anthony Lawrence (2014). AntConc (Version 3.4.3). Tokyo, Japan : Waseda University. http://www.laurenceanthony.net/software

Auger Alain, Barrière Caroline (éds) (2008). « Pattern Based Approaches to Semantic Relation Extraction: a State-of-the-art ». *Terminology*, 14/1.

Aussenac-Gilles Nathalie, Condamines Anne (2007). « Corpus et terminologie ». In : R.T. Pédauque (éd). *La redocumentarisation du monde*, Toulouse : Cepadues, 131-147.

Barsalou Lawrence (2003). « Situated Simulation in the Human Conceptual System ». *Language and Cognitive Processes*, 18, 513-562.

Baumann Klaus-Dieter (2007). « A Communicative-cognitive Approach to Emotion in LSP Communication ». In : Kurshid Ahmad, Margaret Rogers (éds). *Evidence-based LSP*. Berne : Peter Lang, 323-344.

Boleda Gemma (2020). « Distributional Semantics and Linguistic Theory ». *Annual Review of Linguistics*, 6/1, 213-234.

Callies Marcus (2018). « Patterns of Direct Tranzitivation and differences between British and American English ». In : Mark Kaunisto, Mikko Höglund *et alii* (éds). *Changing Structures, Studies in Constructions and Complementation*. Amsterdam/Philadelphia : John Benjamins, 151-167.

Condamines Anne (2002). « Corpus Analysis and Conceptual Relation Patterns ». *Terminology*, 8(1), 141-162.

Condamines Anne (2017). « The Emotional Dimension in Terminological Variation : the Example of Transitivization of the Locative Complement in Fishing ». In : Patrick Drouin, Aline Francoeur *et alii* (éds). *Multiple Perspectives on Terminological Variation*. Amsterdam/ Philadelphia : John Benjamins, 11-30.

Condamines Anne (2018a). « Is "To Fish in a River" Equivalent to "To Fish a River"? A Study at the Crossroads of Cognitive Sociolinguistics and Corpus Linguistics ». *Cognitive Linguistic Studies*, 5/2, 208-229.

Condamines Anne (2018b). « La transitivation des compléments circonstanciels dans le sport et les loisirs, en situation d'implication affective : néologie sémantique ou simple variation argumentale ? ». In : Delphine Bernhard, Maryvonne Boisseau *et alii* (éds). *La néologie en contexte. Cultures, situations, textes*. Limoges : Lambert-Lucas, 217-230.

Condamines Anne (2018c). «Terminological Knowledge Bases ». In : Pedro Fuertes-Olivera : *The Routledge Handbook of lexicography*. London : Routledge, 335-349.

Condamines Anne (2021). « How Can One Explain "Deviant" Linguistic Functioning in Terminology? ». *Terminology*, first published on line https://doi.org/10.1075/term.20029.con.

Condamines Anne, Picton Aurélie (à paraître). « Textual Terminology : Origins, Principles and New Challenges ». In : Marie-Claude L'Homme, Pamela Faber (éds). *Theoretical Approaches to Terminology*. Amsterdam/ Philadelphia : John Benjamins.

Faber Pamela (2012). *A Cognitive Linguistics View of Terminology and Specialized Language*. Berlin : Mouton de Gruyter.

Fabre Cécile, Hathout Nabil, Sajous Franck, Tanguy Ludovic (2014). « Ajuster l'analyse distributionnelle à un corpus spécialisé de petite taille ». In : Actes de la *21e Conférence sur le Traitement Automatique des Langues Naturelles (TALN 2014)*, juin 2014, Marseille, France. Marseille : Université d'Aix Marseille, 266-279.

Firth John Rupert (1957). *Papers in Linguistics 1934–1951*. Oxford : Oxford University Press.

Firth John Rupert (1968). « Linguistic Analysis as a Study of Meaning » In : Frank Robert Palmer (éd). *Selected Papers of J.R. Firth*. Londres : Longman (première édition 1952), 12-26.

Gaudin François (2003). *Socioterminologie – Une approche sociolinguistique de la terminologie*. Bruxelles : De Boeck - Duculot.

Glenberg Arthur, Robertson David (2000). « Symbol Grounding and Meaning : A Comparison of High-Dimensional and Embodied Theories of Meaning ». *Journal of Memory and Language*, 43, 379-401.

Goldberg Adele (1996). « Jackendoff and Construction-based Grammar ». *Cognitive Linguistics*, 7/1, 3-19.

Gries Stefan (2015). « The Role of Quantitative Methods in Cognitive Linguistics : Corpus and Experimental Data on (relative) Frequency and Contingency of Words and Constructions ». In : Jocelyne Daems, Eline Zenner *et alii* (éds). *Change of paradigms - New paradoxes: Recontextualizing Language and Linguistics*. Berlin/ Boston : Walter de Gruyter, 311-325.

Harris Zellig (1954). « Distributional Structure ». *Word*, 10(23), 146-162.

Heylen Kris, Bertels Ann (2016). « Sémantique distributionnelle en linguistique de corpus ». *Langages*, 201/1, 51-64.

Kristiansen Gitte, Dirven René (éds) (2008). *Cognitive Sociolinguistics : Language Variation, Cultural Model, Social Systems*. Berlin : Mouton de Gruyter.

Landauer Thomas (1999). « Latent Semantic Analysis (LSA), a Disembodied Learning Machine, Acquires Human Word Meaning Vicariously from Language Alone ». *Behavioral and Brain Sciences*, 22/4, 624-625.

Lenci Alessandro (2008). « Distributional semantics in linguistics and cognitive research ». *Italian Journal of Linguistics*, 20/1, 1-31.

Marshman Elizabeth, L'Homme Marie-Claude, Surtees Victoria (2008). « Portability of cause-effect relation markers across specialized domains and text genres : A comparative evaluation ». *Corpora*, 3/2, 141-172.

Meyer Ingrid (2001). « Extracting Knowledge-Rich Contexts for Terminography : A Conceptual and Methodological Framework ». In : Didier Bourigault, Marie-Claude L'Homme *et alii* (éds). *Recent Advances in Computational Terminology*. Amsterdam/ New York : John Benjamins Publishing Company, 279-302.

Meyer Ingrid, Bowker Lynne, Eck Karen (1992). « Cogniterm: An Experiment in Building a Terminological Knowledge Base ». In : *Proceedings of 5th EURALEX International Congress on Lexicography*, Tampere, Finland. Tampere : Studia Translatologica, 159-172.

Rastier François (2021). « Data vs corpora ». In : Damon Mayaffre, Laurent Vanni (éds). *L'intelligence artificielle des textes. Des algorithmes à l'interprétation*. Paris : Honoré Champion, 203-246.

Segui Juan (2015). « Évolution du concept de lexique mental ». *Revue de neuropsychologie*, 7/1, 21-26.

Temmerman Rita (2000). *Towards New Ways of Terminological Description. The Sociocognitive Approach*. Amsterdam/ Philadelphia : John Benjamins.

Turney Peter David, Pantel Patrick (2010). « From Frequency to Meaning : Vector Space Models of Semantics ». *Journal of Artificial Intelligence Research*, 37, 141-188.

Wüster Eugen (1976). « La théorie générale de la terminologie - un domaine interdisciplinaire impliquant la linguistique, la logique, l'ontologie, l'informatique et les sciences des objets (Trans.) ». In : H. Dupuis (éd), *Essai de définition de la terminologie. Actes du colloque international de terminologie*. Québec : Régie de la Langue Française, 49-57.