

L'EXTRACTION DE TERMES-CLÉS DE LA PÊCHE À L'AIDE D'OUTILS GNU/LINUX

AGNIESZKA K. KALISKA

Université Adam Mickiewicz, Poznań, Pologne

KALISKA Agnieszka K.: Extracting fishing terminology using GNU/Linux tools. *Jazykovedný časopis (Journal of Linguistics)*, 2021, Vol. 72, No 4, pp. 916 – 926.

Abstract: The technological revolution that has occurred in recent decades has made accessible for researches large textual data collections. At the same time, the development of increasingly sophisticated computer tools provides them with new methods of analyzing texts. In the present study however we examine the functionalities offered by traditional tools, namely GNU/Linux tools, easily accessible via the command line but still unknown among linguists with little or no computer knowledge. Our goal is to show how using the web corpus on the one hand and the processing GNU/Linux tools on the other, we can extract key-terms of fishing jargon.

Key words: web corpus, GNU/Linux tools, key-term, fishing terminology.

1. INTRODUCTION

La révolution technologique qui s'est produite ces dernières décennies a rendu accessibles des collections de données textuelles de grande taille, parmi lesquelles le corpus web occupe une place spéciale. Parallèlement, le développement d'outils informatiques de plus en plus sophistiqués, souvent en accès libre, ouvre aux chercheurs de nouvelles voies d'exploitation des textes. Dans la présente étude nous nous proposons néanmoins d'examiner les fonctionnalités que nous offrent des outils traditionnels, à savoir les outils GNU/Linux accessibles pour tout utilisateur du système Linux via la ligne de commande. Accessibles, certes, ces outils restent cependant peu connus parmi les linguistes n'ayant aucune ou très peu de connaissance de l'informatique. Notre objectif est de montrer comment grâce au corpus web, d'un côté, et au traitement exécuté à l'aide d'outils GNU/Linux de l'autre, nous pouvons approfondir notre connaissance du vocabulaire spécialisé. Dans cette recherche nous utiliserons le corpus OSCAR (Open Super-large Crawled ALMAnaCH coRpus), le plus grand corpus multilingue disponible librement (Ortiz – Sagot – Romary, 2019).

Par *pêche* nous entendons la pêche *récréative*, dite aussi *pêche de loisir*, *récréationnelle* ou *plaisancière*. Nous commencerons par une brève caractéristique du jargon de la pêche et des méthodes de collecte de données lexicales, à commencer par des enquêtes traditionnelles, une recherche sur le terrain, jusqu'au traitement des données textuelles de grande taille. L'étape suivante de notre étude consistera à présenter les données lexicales collectées. Nous finirons par quelques conclusions concernant les avantages et les inconvénients de la méthode employée.

2. CARACTÉRISTIQUE DU JARGON DE LA PÊCHE

Nous nous soucrivons à la définition du *jargon* proposée par D. François-Geiger (1988) selon qui il s'agit d'« un parler technique qui peut être ésotérique pour le profane, mais dont la fin n'est pas de masquer l'objet du discours ; elle est, au contraire, d'en rendre l'expression plus rigoureuse, plus spécifique ». Les mots de la pêche servent ainsi à *nommer* – de manière juste, brève et précise (cf. aussi : Sourdot, 2011, p. 1) – les outils et méthodes employés pour attraper le poisson.

La deuxième fonction, non moins importante, est *communicative*. Déjà le fait que *pêcheur* ne désigne dans le cadre de la présente recherche ni un métier ni une formation diplômante (car il s'agit d'une activité récréative, non commerciale) rend la pêche accessible pour tous à toute étape de la vie. Le *savoir-pêcher* se transmet ainsi de manière informelle, d'une génération à l'autre, en famille et entre amis, quelque soient l'âge, le sexe, l'expérience professionnelle ou encore la situation sociale de nouveaux amateurs. Dans ces échanges des forums de discussion et des blogs sont une source d'informations précieuse, permettant à un débutant de retracer le parcours nécessaire pour atteindre son objectif. Ainsi le jargon de la pêche permet-il non seulement de transmettre les connaissances mais aussi d'entrer en contact, de dialoguer avec les autres et de nouer des liens durables. Il assume alors la troisième fonction – ie. *identitaire* – car l'échange est d'importance cruciale pour assurer le lien social.

Le dernier trait caractéristique du vocabulaire de la pêche réside dans la dimension ludique de celui-ci (cf. Sourdot, 2018 ; Kaliska, 2018). C'est sur le web qui est un lieu d'échange libre de connaissances que nous pouvons observer l'expressivité et la spontanéité du jargon de la pêche, et très probablement les premières traces écrites de désignations futures. Il ne s'agit donc pas d'un strict technoclecte. Notons ce qu'annoncent les en-têtes de différents sites privés où on lit que la pêche est « un art » ou « une façon de vivre ». Ils témoignent qu'il s'agit pour leurs auteurs d'un véritable dévouement qui trouvera son reflet dans la langue utilisée : pleine d'humour, imagée et technique à la fois.

3. MÉTHODES DE COLLECTE DE DONNÉES LEXICALES

Pour un linguiste non pêcheur l'accès à ce vocabulaire n'est possible que par une des voies suivantes : la première étant un travail de terrain peut consister à distribuer, collecter et interpréter les enquêtes, ce qui est une méthode de travail plutôt coûteuse et insuffisante car, au final, les entretiens, voire les observations directes, permettant de résoudre des questions complexes peuvent s'avérer indispensables. L'efficacité de ces derniers réside justement dans la possibilité d'obtenir une image authentique du domaine dans son environnement naturel, sans que cela signifie pour autant qu'on n'aura pas de difficultés épistémologiques ou méthodologiques à surmonter.¹ Deux méthodes suivantes

¹ V. Labov, 1972 (pour le paradoxe de l'observateur) et Blanchet, 2012 (pour l'observation participante).

consisteraient à explorer soit les éditions papier (p.ex. : la presse de spécialité) soit le web (p.ex. : les blogs). Ici, comme pour les enquêtes, les observations de près des actions et des interactions *in situ* peuvent aider à comprendre mieux le matériel analysé.

Les méthodes de collecte de données citées ne s'excluent pas et peuvent être complémentaires. La question se pose néanmoins quel rôle attribuer à ces différentes sources ? Illustratif ou heuristique ? La question est d'autant plus complexe, si l'on se limite – comme c'est le cas de notre étude – à la dernière des sources citées, notamment le web, c.-à-d. l'ensemble des productions linguistiques sous forme de documents numériques librement consultables de chez soi par le biais d'un moteur de recherche, *en ligne*. Or, ce même ensemble devient – grâce aux projets tels que Common Crawl ou OSCAR (Ortiz – Sagot – Romary, 2019) – accessible aussi en forme brute (*plain text*).²

En même temps, le développement récent d'outils informatiques mis à la disposition de linguistes leur a permis d'accélérer dans leur travail : trouver et calculer des (co-) occurrences de mots sont-ils devenus plus faciles à exécuter dès que des outils spécifiquement développés pour ce type de tâches sont devenus disponibles (p.ex. : MeWeX, LEM et d'autres outils élaborés pour l'analyse des textes polonais par le consortium Clarin-PL, cf. p.ex. : Piasecki, 2014 ; cf. aussi : Drouin, 2015). Dans la présente étude nous nous proposons néanmoins d'examiner les fonctionnalités que nous offrent les outils traditionnels, à savoir les outils GNU/Linux, accessibles pour tout utilisateur du système Linux via la ligne de commande, tels que *cat* (et ses variantes, p.ex. : *zcat*), *grep* (et ses variantes, p.ex. : *egrep*), *sed*, *sort*, *tr*, *uniq*, *wc* (cf. p.ex. : Shotts, 2012), auxquelles s'ajoutent les expressions dites *rationnelles* (parfois appelées *normales*, *régulières* ou, tout court, les *motifs*, dites encore les *regex* par la composition des mots anglais *regular* et *expression*). Les expressions régulières sont un outil très puissant nous permettant d'accéder à des ensembles de chaînes de caractères (séquences de texte) prédéfinies selon une connaissance préalable de structures dont on abstrait les caractéristiques les plus saillantes pour construire le corps du motif. À l'aide des programmes en ligne de commande nous pouvons ensuite extraire ce qui nous intéresse, le trier, dédupliquer, modifier, compter et numéroter, de manière à obtenir des résultats que l'on peut ensuite soumettre au dépouillement terminologique manuel.

4. WEB EN TANT QUE CORPUS

Le corpus web est parfois défini par la formule de trois *V* : *volume*, *vélocité* et *variété*. Ses données sont massives, hétérogènes (textes, images et sons) et se renouvellent en permanence (cf. Longhi, 2017 ; cf. aussi : Fouqueré – Isaac, 2003, p. 115 pour *volatilité* de données). La collecte automatisée de ces différentes données produites massivement est possible grâce aux pratiques de crawling (indexation des pages web) et de scraping (extraction d'informations issues d'Internet). Telle est, notons-le, la vocation de l'organisation Common Crawl : explorer les pages web et

² Un tel corpus fourni sans métadonnées est, en principe, destiné à être utilisé dans le TAL.

fournir au public, gratuitement, ses archives sous forme d'ensembles de textes. Le corpus OSCAR, coordonné par l'équipe ALMAAnaCH dans le centre de recherche Inria de Paris, est un ensemble de textes multilingue obtenu par filtrage du corpus Common Crawl ; il est à télécharger librement sur le site du projet.

Les données issues d'Internet sont ainsi stockées sous forme de fichiers texte brut. La première caractéristique d'un tel ensemble est qu'il est gros, varié et bruité. On y trouvera non seulement toute une variété de textes et de registres mais aussi des pages entières de données qui pour un linguiste lexicographe sont plutôt de moindre qualité. Ce seront, par exemple, des énumérations de chiffres et de signes spéciaux qui, dans un ensemble de textes battus (*shuffled*), ne disent pratiquement rien.

D'autres caractéristiques peuvent découler de la spécificité du discours électronique médié qui a des traits à la fois oraux et écrits. Y sont présents à différents et pas toujours les mêmes degrés (p.ex. chat vs. blog ; cf. Panckhurst, 2007) des phénomènes comme écritures hors norme, absence ou emploi excessif de certaines formes grammaticales (cf. p.ex.: Amitay, 1999, cité après Fouqueré – Issac, 2003, p. 116), fautes de frappe, omission spontanée ou systématique de signes diacritiques, fautes d'orthographe dues à l'homophonie des formes ou à l'écriture phonétique.

Les traits cités posent des difficultés de traitement, certes, mais il paraît que la plus grande difficulté consiste à repérer dans un corpus aussi immense qu'OSCAR (23,206,776,649 mots graphiques) les structures utiles pour la recherche. Pour ce faire, il faut avoir une connaissance préalable de mots clés et de structures potentiellement intéressantes et sur cette base construire un ou des motifs utiles pour la recherche plus approfondie de termes-clés (termes-candidats).

5. DÉPOUILLEMENT TERMINOLOGIQUE

5.1 Sous-corpus de la pêche

La constitution d'un corpus cohérent à partir de données aussi variées que le web n'étant pas facile, nous nous sommes servis d'un mot clé *pêche* afin de construire sur sa base un motif – à savoir : `(^|\s)\b(p|P)êch\w+\b` – qui nous a permis d'accéder à 1,312,157 lignes contenant la séquence *pêch(e)-* (en début du mot pour distinguer entre *pêche*, *pêcheur* et *pêcher...*, d'un côté, et *empêcher* et *dépêcher...* de l'autre), ce qui constitue env. 0,28% du corpus entier (461,343,098 lignes).³

³ Tout fichier texte comporte un certain nombre de caractères, à savoir : caractères imprimables, espaces et sauts de ligne. Ces derniers sont des caractères de contrôle indiquant le passage à la ligne suivante (*line feed*). Les limites d'une ligne ne sont donc pas ce qu'on voit aux frontières d'une fenêtre d'affichage. Dans notre sous-corpus français, la longueur moyenne d'une ligne égale à 656 caractères (sans compter les sauts de ligne), tandis que la ligne la plus longue s'y compose de 950,890 caractères et la plus courte en a une centaine. Encore que les lignes doubles aient été éliminées avant que le corpus eût été mis en ligne par ses créateurs, certaines se répètent – ce sont des lignes qui ne diffèrent, par exemple, que d'un caractère ou d'un mot, il fut donc impossible de les détecter à l'étape du filtrage.

Bien que le motif que nous avons proposé nous ait permis d’accéder relative- ment vite à un nombre considérable d’emplois, on n’a pas pu à cette étape-là contrer les difficultés qu’occasionne l’homonymie lexicale. En effet, il y a *pêche* qui signifie ‘action ou manière de pêcher’ (cf. TLFi) et *pêche*, son homographe, qui signifie ‘fruit du pêcher’. Analogiquement, *spinning* désigne une technique de pêche ou une activité sportive de cardio-training. Ce type de difficultés ne peuvent être résolues autrement qu’au cours d’une lecture minutieuse des exemples extraits.

5.2 Extraction des termes-clés : expressions régulières

Les expressions régulières facilitent considérablement la fouille du corpus. Nous nous en sommes servis pour repérer, à l’aide de la commande `egrep`, des noms de méthodes de pêche. Les motifs ont été bâtis à la base de structures que nous avons considé- rées comme potentiellement utiles pour le repérage des termes recherchés (p.ex.: *X est une méthode*). Notons que le même mot peut apparaître dans différentes structures :

Structure > Motif > Exemples			
(\ble\b \bla\b)\s\b\w+\b(\s\b\w+\b){,2}\sest\s(une la)\s(méthode technique)			
Ex. : [L]a pêche à la dandinette est une technique à utiliser lorsqu'on a repéré un groupe de dorés.	la bolognaise la bouée la bouillette délavée la carpe à la batterie la carpe au coup la carpe au pellet waggler la charfia ou chrafi la cuiller la dandine la dandinette la ligne à la traine	la madrague la mouche au tenkara la mouche en nymphe la mouche noyée la mouche sèche la pêche de fond la pêche en float tube la pêche en nymphe la pêche en texan la pêche verticale le down shot le drop shot	le feeder à longue distance le filet barrière le filet maillant le fireball le jig le jiggling lourd le tenkara le toc le vif le wacky
\bla\b\spêche\s(à au)\s\b\w+\b(\s\b\w+\b){,2}			
Ex. : Idéale pour la pêche à la dandinette, cette canne devient inexploitable au lancer.	la pêche à la bombette la pêche à la carpe la pêche à la cuillère la pêche à la dandinette la pêche à la dérive la pêche à la ligne morte la pêche à la mouche noyée la pêche à la mouche sèche la pêche à la nymphe la pêche à la pirouette la pêche à la traîne la pêche à la verticale la pêche à la volée la pêche à rôder	la pêche à soutenir la pêche à vue la pêche au boobie la pêche au booby la pêche au buscle la pêche au coup la pêche au fire la pêche au feeder la pêche au flotteur la pêche au jig la pêche au lancer la pêche au léger la pêche au leurre lancé la pêche au leurre manié	la pêche au mort manié la pêche au mort posé la pêche au posé la pêche au poser la pêche au quiver la pêche au rubber jig la pêche au shad en linéaire la pêche au stream la pêche au streamer la pêche au swing tip la pêche au tenya la pêche au toc la pêche au vif

\bpêche\b\s\w+ (ante elle que ère enne one ane ée euse eure) (\s \n §)			
Ex. : Pour la pêche itinérante le panier se porte en bandoulière.	la pêche carnassière la pêche civelière la pêche civelse la pêche coquillière la pêche germonière	la pêche harengière la pêche itinérante la pêche journalière la pêche morutière la pêche maquereautière	la pêche piroguière la pêche planante la pêche supersonique la pêche télescopique la pêche thonnière

Tab. 1. Extraction des noms de méthodes de pêche.

Certains syntagmes que nous avons pu extraire renvoient à la pêche non récréative, p.ex.: *pêche baleinière*, *pêche électrique*, *pêche halieutique*, *pêche haussière*, *pêche hauturière* (dite aussi *pêche au large*), *pêche nourricière*.

Certaines appellations sont des mots anglais (p.ex.: *surfcasting*, *drop shot*, *jig*). Celles-là et d'autres importations anglaises sont, en effet, assez fréquentes dans le jargon de la pêche, ce qui peut s'expliquer par le fait que les Anglais ont été les premiers à faire de la pêche une activité récréative (Sourdou, 2018, p. 276). Grâce aux motifs faisant référence à la structure interne de certains dérivés (p.ex.: *-ing*), nous avons pu extraire des noms de méthodes (p.ex.: *popping*, *spinning*, *trolling*), d'activités (p.ex.: *cofishing*) et de qualités (employés comme épithètes auprès des noms de leurres, p.ex.: *sinking* et *suspending* – en parlant de woblers). Pareillement pour les composés avec *bait* (fr. 'leurre'). Ces derniers n'ont pas, en général, d'orthographe stable, ce que confirment les exemples de corpus : ils s'écrivent soudés ou séparés par un espace, ou encore avec un tiret d'union – cf. Tableau 2.

5.3 Extraction des termes-clés : *n*-grammes

Dans le traitement automatique des langues, les *n*-grammes renvoient aux suites de *n* caractères (ce à quoi on recourt couramment dans l'apprentissage automatique, ang. *machine learning*) ou aux suites de *n* mots. Dans le cas des *bi*-grammes, le *n* égale à 2, dans le cas des *tri*-grammes, le *n* égale à 3, et ainsi de suite. L'idée est qu'à partir d'une séquence de *n* caractères ou de *n* mots il est possible d'évaluer la probabilité d'occurrence pour un caractère ou un mot suivant. Les méthodes statistiques sont, en effet, en usage également pour évaluer le potentiel terminologique des mots, encore que depuis un certain temps elles cèdent la place aux réseaux neuronaux.

Structure > Motif > Exemples			
\b\w+ing\b			
Ex. : Toutes les techniques sont praticables ici mais nous privilégions la pêche aux leurres (lancer et jigging).	baitcasting casting cofishing cranking fluorocasting flyfishing jigging matching	pitching popping powerfishing rockfishing rolling slow-jigging spinning	sportfishing stalking streetfishing surfcasting trolling twitching urban fishing

\b\w+(- \s)?bait\b		
Ex. : Le crankbait peut également être coulant ou suspending permettant d'accéder aux couches d'eaux inférieures.	bigbait (big-bait, big bait)	hookbait (hook bait, hook-bait)
	buzzbait (buzz bait, buzz-bait)	jerkbait
	chatterbait (chatter bait)	powerbait (power bait)
	cran(k)bait (cran(k)-bait, cran(k) bait)	propbait (prop bait)
	glidebait (glide bait)	softbait (soft bait)
	groundbait	spinnerbait (spinner bait,spinner-bait)
	hardbait (hard bait, hard-bait)	stickbait (stick bait)

Tab. 2. Extraction des importations anglaises en *-ing* et *bait*.

Or, les *n*-grammes peuvent servir de point de départ pour l'extraction des termes-clés syntagmes. Il ne suffit pas, certes, que les mots qui se suivent dans le texte soient deux, trois ou quatre, pour constituer un syntagme (p.ex. : *pêche à, pêche à la* et *la pêche à la* sont des *n*-grammes sans pour autant être syntagmes). Dans la suite nous expliquons les étapes de l'extraction de *n*-grammes avant qu'ils soient soumis à une lecture permettant d'en tirer ceux qui sont, en fait, des termes de pêche.

5.3.1 Termes-candidats avec les mots clés : *canne, flotteur, leurre et moulinet*

Ci-dessous nous présentons quelques exemples de termes-clés qui furent extraits à partir des *n*-grammes à deux, trois et quatre mots textuels.⁴ Nous nous limitons ci-dessous aux syntagmes dans lesquels les substantifs *canne, moulinet* et *leurre* sont des noyaux.

Une première série d'exemples a pour le noyau le substantif *leurre* ('appât artificiel utilisé pour la capture des poissons' – cf. TLFi). Une fois les *bi*-grammes *leurre X* extraits, nous les avons passés au crible de la sélection automatique : nous avons commencé par l'élimination des mots vides (*stop words*) tels que les articles, les prépositions, les pronoms et – puisque nous avons décidé de nous limiter dans cet exercice aux syntagmes nominaux – nous avons passé à l'élimination des formes conjuguées de verbes (p.ex. : *a, est, dépend*). Nous avons ainsi obtenu env. 700 emplois uniques que nous avons soumis à un dépouillement manuel. Notons que déjà pour les *tri*-grammes la procédure était différente. Étant donné que trois mots peuvent correspondre à une synapsie et que les synapsies sont fréquemment utilisées dans les

⁴ Voici le haut de la liste de 10,810,023 *bi*-grammes : *de la* (1,248,804 occurrences), *de l'* (764,699), *à la* (525,833), *la pêche* (381,253), *de pêche* (339,510). Notons qu'en haut de cette liste fréquentielle de séquences pour la plupart des cas inutiles (car ni syntagmes ni termes recherchés) le mot *pêche* apparaît dans deux séquences différentes (non chevauchantes). Les *tri*-grammes, *canne à pêche* et *bateau de pêche* occupent les positions soixante-onzième et cent-seizième sur la liste de 28,237,535 séquences, avec 10,291 et 8,147 occurrences (sans compter les variantes mal écrites, p.ex. : *peche*, ou grammaticales, p.ex. : *bateaux de pêche*), tandis que les *quadri*-grammes *pêche à la mouche* et *pêche à la ligne* se trouvent déjà à la position quatrième et à la position septième du haut de la liste fréquentielle de 39,101,958 suites avec, respectivement, 12,363 et 10,492 occurrences. Ces données confirment la spécificité sémantique du corpus étudié.

nomenclatures techniques, ce que Benveniste constata déjà quand il introduisait la notion de synapsie en linguistique il y a 50 ans (Benveniste, 1966, p. 172), l'élimination des mots vides ne pouvait plus se faire sans avoir soumis ceux-là à une sélection, tenu compte de la position centrale qu'un mot *vide*, tel la préposition *à* ou la préposition *de*, occupe dans le composé syntactique (cf. Tableau 3).

<i>n</i> -grammes	Exemples		
<i>bi</i> -grammes : <i>leurre X</i> (syntagme nominal) Ex. : <i>Le leurre lipless ou lipless crankbait est un leurre coulant dépourvu de bavette.</i>	leurre anguille leurre artisanal leurre artificiel leurre brochet leurre casting leurre dur leurre écrevisse leurre fluo	leurre fouetté leurre grenouille leurre jig leurre lesté leurre libellule leurre lipless leurre lézard leurre maison	leurre ondulant leurre pencil leurre plombé leurre plongeant leurre souple leurre tournoyant leurre vairon leurre worm
<i>tri</i> -grammes : <i>leurre Prép N</i> (synapsie) Ex. : <i>Ce leurre de suspension combine le profile d'un poisson appât et l'action d'un poisson blessé.</i>	leurre à bavette leurre à brochet leurre à dandiner leurre à distance leurre à flapper leurre à hameçon	leurre à hélice leurre à jigger leurre à perche leurre à truite leurre à truites leurre de pêche	leurre de prospection leurre de spinning leurre de surface leurre de suspension leurre de traîne leurre de traque
<i>quadri</i> -grammes : <i>leurre Prép (N Adj N N Adj N)</i> (synapsie) Ex. : <i>La cuiller tournante (par opposition à cuiller ondulante) est un leurre à palette tournante.</i>	leurre à action moyenne leurre à collerette double leurre à double caudale leurre à hameçon simple leurre à la bombette leurre à palette tournante leurre à tête plombée leurre de grande prospection leurre de pêche wobbler leurre de pêche crankbait leurre de pêche cuillère leurre de pêche modulaire	leurre de pêche spinnerbait leurre de pêche profonde leurre de pêche télescopique leurre de pêche vibrant leurre de pêche swimbait leurre de pêche worm leurre de type finesse leurre de type crankbait leurre de type jerkbait leurre de type mouche leurre de type shad	

Tab. 3. Termes-clés avec *leurre* comme noyau du syntagme

Parmi les *quadri*-grammes extraits certains sont des syntagmes nominaux libres, non terminologiques, dont les satellites fournissent des caractéristiques à propos des traits en quelque sorte accidentels de leurres tels que leur couleur (p.ex.: *leurre de couleur criarde, leurre de couleur discrète, leurre de couleur sombre, leurre de couleur jaune*), poids (p.ex.: *leurre de faible grammage*), forme (p.ex.: *leurre de forme longiligne, leurre de forme triangulaire*) et taille (p.ex.: *leurre de grande taille, leurre de grosse taille, leurre de petite taille, leurre de taille moyenne*).

Trois séries d'exemples qui suivent ont pour les noyaux les mots clés *canne, flotteur* et *moulinet* :

<i>n</i> -grammes – Exemples		
<i>bi</i> -grammes : (<i>canne</i> <i>flotteur</i> <i>moulinet</i>) (<i>N</i> <i>Adj</i>) (syntagme nominal)	<i>tri</i> -grammes : (<i>canne</i> <i>flotteur</i> <i>moulinet</i>) <i>Prép N</i> (synapsie)	<i>quadri</i> -grammes : (<i>canne</i> <i>flotteur</i> <i>moulinet</i>) <i>Prép (N Adj N N Adj N)</i> (synapsie)
canne anglaise canne baitcasting canne casting canne feeder canne silure canne spinning canne toc canne téléréglable canne télescopique	canne au posé canne à appâts canne à déboité canne à emboîtement canne à emmanchement canne à pêche canne de jig canne de spinning canne de traîne	canne à pêche anglaise canne à pêche blanche canne à pêche bolognaise canne à pêche jigging canne à pêche spinning canne à pêche surfcasting canne à pêche télescopique canne à scion souple canne de surfcasting léger
flotteur boule flotteur carotte flotteur glisseur flotteur lumineux flotteur poire flotteur sondeur flotteur tige	flotteur de cassant flotteur de pêche flotteur de sonde flotteur goutte d'eau flotteur pêche mer flotteur pour carpes flotteur sans antenne	flotteur de forme crayon flotteur de forme poire flotteur en forme de stick flotteur de pêche ascendant flotteur de pêche coulissant flotteur de pêche balle flotteur de pêche nuit
moulinet amovible moulinet baitfinesse moulinet débrayable moulinet casting moulinet cranking moulinet match moulinet spinncasting moulinet spinning moulinet verrouillable	moulinet de pêche moulinet de surfcasting moulinet frein avant moulinet frein arrière moulinet grande capacité moulinet jigger électronique moulinet lancer lourd moulinet large arbor moulinet tambour tournant	moulinet à bobine fixe moulinet à bobine libre moulinet à frein avant moulinet à grande récupération moulinet à ratio lent moulinet à tambour tournant moulinet à tambour fixe moulinet de pêche centerpin moulinet de pêche rotatif

Tab. 4. Termes-clés avec *canne*, *flotteur* et *moulinet* comme noyaux des syntagmes.

À part les exemples cités, d'autres composés synaptiques furent extraits et soumis à la vérification manuelle, p.ex.: *bobine à tambour fixe*, *cuillère ondulante*, *cuillère tournante*, *cuillère vaironnée*, *émerillon à agrafe*, *émerillon à billes*, *émerillon baril*, *émerillon rolling*, *hameçon à œillet*, *œillet de guidage*, etc.

6. CONCLUSION

La brève présentation que nous venons de faire montre à quel point les outils GNU/Linux, et les expressions régulières, peuvent s'avérer utiles dans la collecte des candidats-termes nominaux. La connaissance des propriétés morphosyntaxiques des syntagmes, composés réguliers et synaptiques, permet au linguiste d'accéder relativement vite aux séquences qu'il peut ensuite soumettre à une vérification manuelle et en sélectionner celles pour lesquelles le potentiel terminologique n'est pas uniquement l'affaire de fréquence, surtout lorsqu'on travaille sur un corpus aussi

immense et bruité que le corpus web et que l'objet de notre recherche – en l'occurrence le jargon de la pêche – n'y est qu'un pour mille. C'est d'ailleurs pour cette raison que le corpus traité a, certes, rendu possible « la quête de l'inconnu », il « a montré », comme disait Scheer (2004), sans pour autant « démontrer » la totalité du parler étudié. Et pour qu'une vision ainsi bâtie soit plus complète, il faudrait – à part l'extraction des termes nominaux – prévoir un traitement qui aboutirait à la sélection des verbes et des constructions verbales typiques pour le jargon de la pêche, telles que *prospector les zones profondes*, *scanner le fond* (en parlant des humains), et *se jerker*, *travailler* (en parlant des leurres).

Bibliographie

AMITAY, Einat : Anchors in context: A corpus analysis of web pages authoring conventions. In : Words on the Web – Computer Mediated Communication. Eds. L. Pemberton – S. Shurville. Éd. Intellect Books 1999.

BLANCHET, Philippe : La linguistique de terrain. Méthode et théorie. Rennes : Presses universitaires de Rennes 2012.

DROUIN, Patrick : Acquisition automatique de termes : simuler le travail du terminologue. In : Études de linguistique appliquée, 2015, Vol. 180, No 4, pp. 417–427.

FOUQUERÉ, Christophe – ISSAC, Fabrice : Corpus issus du Web : constitution et analyse informationnelle. In : Revue québécoise de linguistique, 2003, Vol. 32, No 1, pp. 111–134.

FRANÇOIS-GEIGER, Denise : 1988, Les paradoxes des argots. In : Actes du Colloque culture et pauvreté. Eds. A. Lion – P. de Meca. 1988, pp. 17–24, Tourette 13-15 décembre 1985, La Documentation française.

GANASCIA Jean-Gabriel : Les *big data* dans les humanités. In : Critique, 2015, No 819-820, pp. 627–636.

KALISKA, Agnieszka : Potoczność a terminologiczność socjolektu wędkarskiego. Polskie i francuskie nazwy technik wędkarskich. In : Orbis Linguarum, Vol. 49, Dresden-Wrocław : Neisse Verlag & Oficyna Wydawnicza ATUT 2018, pp. 109–126.

LABOV, William : Some Principles of Linguistic Methodology. In : Language in Society, 1972, Vol. 1, No 1, pp. 97–120.

LONGHI, Julien : Humanités, numérique : des corpus au sens, du sens aux corpus. In : Questions de communication, 2017, Vol. 31, pp. 7–17.

ORTIZ, Pedro J. – SAGOT, Benoît – ROMARY, Laurent : OSCAR – Open Super-large Crawled ALMAnaCH coRpus. Disponible sur : <https://oscar-corpus.com/ux> normes (accès : 12-14 octobre 2020).

PANCKHURST, Rachel : Discours électronique médié : quelle évolution depuis une décennie ? In : La langue du cyberspace : de la diversité aux normes. Ed. J. Gerbault. Paris : L'Harmattan 2007, pp. 121–136.

PIASECKI, Maciej : User-driven language technology infrastructure – the case of Clarin-PL. In : Proceedings of the 9th Language Technologies Conference. Ljubljana : Information Society 2014.

SCHEER, Tobias : Le corpus heuristique : un outil qui montre mais ne démontre pas. In : Corpus, 2004, No 3, pp. 153–192.

SHOTTS, William E., Jr. : The Linux Command Line. Éd. No Starch Press 2012.

SOURDOT, Marc : Les mots de la pêche ou « emprunter c'est jouer ». In : Journée d'hommage à A.-M. Houdebine. La Rochelle 17 juin 2011. Disponible sur : l.brunethunault.free.fr/17juin/marcSourdot.pdf.

SOURDOT, Marc : Il faut appeler un « shad » un shad. In : Le poids des mots. Hommage à Alicja Kacprzak. Eds. A. Konowska – A. Woch – A. Napieralski – A. Bobińska. Łódź : Wydawnictwo Uniwersytetu Łódzkiego 2018, pp. 275–281.

Sitographie

Mots vides français en ligne. Disponible sur : <https://countwordsfree.com/stopwords/french>
Ressources lexicales sur le site d'Yann van der Cruyssen. Disponible sur : <http://www.nurykabe.com>

TLFi : Trésor de la langue Française informatisé. Disponible sur : <http://www.atilf.fr/tlfi>,
ATILF - CNRS & Université de Lorraine.

Remerciements :

Je tiens à remercier Karol Kaczmarek, doctorant en informatique à l'Université Adam Mickiewicz/Applica.AI, pour son aide dans le téléchargement et le filtrage du corpus. Un grand merci aussi à tous les internautes, souvent anonymes, qui contribuent éminemment à l'échange de connaissances liées au traitement automatique du texte.