Expériences avec d'autres logiciels ANA et TermoStat

Dans le chapitre précédent, nous avons présenté le logiciel Unitex et illustré quelques pistes pour des travaux en terminologie. Dans ce chapitre, nous relatons en revanche quelques expériences menées sur d'autres logiciels, basés sur d'autres types de méthodes (statistiques et hybrides). La recherche a privilégié des logiciels conçus en vue d'applications terminologiques, comme ANA et TermoStat. Nous avons cru que cela pouvait être intéressant, d'autant plus qu'Unitex n'est pas un logiciel conçu expressément pour la terminologie. Pendant notre parcours de thèse, nous avons aussi pu tester un analyseur syntaxique opérationnel, SYNTEX, qui n'a pas été non plus conçu en vue d'applications terminologiques, mais qui a été développé à partir des travaux sur le logiciel de structuration de terminologie LEXTER (BOURIGAULT 1994). Toutefois, nous nous limitons à reporter ici les expériences avec les deux premiers logiciels cités[143].

Avant de passer à l'analyse de ces expériences, nous détaillons les quelques difficultés qui ont accompagné ces tentatives, explicables surtout par l'indisponibilité de certains logiciels sur le marché et par la nécessité de posséder des compétences poussées en informatique.

2.1. Tentatives d'essai de quelques logiciels : difficultés rencontrées

On se souviendra que nous avons dressé un panorama d'outils dans I.3.5. et sous-sections relatives : 3.5.1., 3.5.2. et 3.5.3.). Aux logiciels décrits, il faut en ajouter quelques autres que nous avons recherchés et qui sont désormais indisponibles, en raison de la rapidité avec laquelle ces outils évoluent. Tel a été le cas de Terminology Extractor, outil pour l'extraction terminologique travaillant sur l'anglais et le français, développé par Etienne Cornu pour l'entreprise Chamblon Systems Inc. Cambridge (Ontario, Canada). Comme le téléchargement n'a pas donné de résultats, nous avons contacté l'auteur, qui nous a informée de l'indisponibilité de Terminology Extractor. La même situation s'est reproduite pour l'outil Mantex, conçu pour le système d'exploitation Macintosh en 2000 par P. Frath. Cet outil était fondé sur des techniques statistiques et visait l'identification des syntagmes répétés dans un corpus.

Nous avons également essayé de tester LEXTER de D. Bourigault (1993), mais comme la propriété du logiciel est désormais détenue par Électricité De France (EDF) cela n'a pas été possible.

Dans d'autres cas, les obstacles ont été l'insuffisance des manuels d'installation et d'utilisation de ces logiciels, limités à un fichier « Read me » de deux pages et évidemment adressés à des professionnels chevronnés : nous nous référons aux outils FASTR de Ch. Jacquemin et ACABIT de B. Daille. En ce qui concerne ACABIT, nous avons rencontré un empêchement supplémentaire : le système n'accepte que des données prétraitées. Ce qui

[143] Dans un premier temps, la tentative de tester TermoStat s'était révélée infructueuse. Nous avons donc entrepris de tester le logiciel SYNTEX comme exemple de traitement de corpus par un logiciel basé sur des méthodes hybrides. Lorsqu'une nouvelle tentative sur TermoStat a réussi, nous avons préféré garder ce dernier comme représentant de la catégorie des approches hybrides. En tout cas, nous remercions vivement Didier Bourigault pour sa disponibilité à l'occasion du test de SYNTEX.

implique le recours à des programmes extérieurs au logiciel et par conséquent des temps plus longs pour l'obtention des résultats[144].

2.2. Expérience d'extraction terminologique par le logiciel ANA

Le premier logiciel que nous avons pu tester a été ANA (Apprentissage Naturel Automatique), développé par Chantal Enguehard entre 1992 et 1995[145]. Malgré son indisponibilité sur le marché, il est toutefois possible de tester ANA en contactant par mail l'équipe de Chantal Enguehard, qui continue à travailler sur cet outil. Nous rappelons qu'ANA est un logiciel basé sur deux méthodes algorithmiques, qui ne recourt à aucune ressource linguistique pour le traitement des textes : il s'agit d'un logiciel fondé uniquement sur des critères statistiques.

Nous avons contacté Chantal Enguehard en septembre 2010, pour lui demander de tester ANA. Pour ce faire, nous avons proposé de traiter deux textes : un premier texte, Alger B_019, de petite taille (environ 1 500 mots), qui n'a pas pu être traité en raison de sa petite taille, ce qui confirme un point faible des approches statistiques ; un second texte, Corpus PTC, de taille légèrement inférieure à notre corpus d'étude actuel (environ 164 000 mots, pour un écart d'environ 15 000 mots)[146]. Les résultats de l'extraction des termes faite par ANA sur ce corpus textuel nous ont été fournis par courriel sous forme d'un tableau de texte. Ces résultats sont organisés par ordre alphabétique dans trois colonnes : dans une première, les candidats termes extraits ; dans une deuxième, le nombre d'occurrences ; dans la troisième, les segments de texte d'où le candidat terme a été extrait avec le nombre d'occurrences pour chaque segment :

Cure de boisson	1	(cure de boisson, 30) (cures de boisson, 1)

Figure 1 : exemple du tableau des résultats de l'extraction par ANA.

Les candidats termes extraits par ANA sont au nombre de 2085. Aucune différence n'est faite entre termes simples et complexes, tous les candidats termes font l'objet du même fichier. L'auteure nous a informée que le seuil minimal de fréquence établi pour l'extraction a été de 3 occurrences et que le nombre de termes présents dans le bootstrap était de 7, mais nous ne savons pas quels étaient ces termes.

2.2.1. Résultats de l'expérience avec ANA : rappel et précision

L'évaluation de ces résultats a été faite en termes de rappel et de précision, suivant les pratiques courantes d'évaluation d'expériences de ce genre. Par *rappel* nous entendons

[144] En outre, ces deux logiciels ne peuvent être utilisés qu'avec le système d'exploitation Linux. La maîtrise de Linux n'est pas encore très répandue dans les milieux des sciences humaines.

[145] Nous renvoyons à II.1.5.1. pour la description de ANA. Il est vrai que le logiciel est désormais quelque peu daté, mais il s'agit du seul extracteur de terminologie à base d'approches statistiques que nous ayons pu essayer ; nous en remercions vivement l'auteure.

[146] Il s'agit d'une version presque définitive du corpus d'étude, dont l'écart avec la version définitive ne pose pas de problèmes pour ce genre d'expérience.

le pourcentage de termes pertinents extraits par ANA par rapport aux termes manuellement identifiés comme pertinents dans le fichier soumis à l'analyse (corpus PTC). Par *précision*, en revanche, nous entendons les termes pertinents sur la totalité des termes extraits.

Pour le calcul du rappel, dans le fichier de départ les termes ont été isolés manuellement à l'aide de balises, comme dans l'exemple suivant :

<terme>phénomène de Raynaud</terme>.

Comme on peut l'imaginer, l'annotation manuelle de tous les termes du corpus soumis à l'analyse aurait requis beaucoup de temps. Nous avons donc mené le calcul du rappel sur une portion du corpus, réunissant des textes variés et dont la taille atteint environ 26 000 mots (16% du corpus). Il s'ensuit que le taux de rappel que nous reportons est un taux approximatif.

Les outils linguistiques de support à cette phase d'identification des termes présents dans la portion de corpus retenue pour le rappel ont été le *GDT* (*Grand Dictionnaire Terminologique*) et le *TLFi* (*Trésor de la Langue Française informatisé*). Ces mêmes outils nous ont servi lors du calcul de la précision.

1 504 termes ont été identifiés dans la portion de corpus choisie pour l'évaluation du rappel. Sur ces 1 504 termes, 527 figurent dans la liste des candidats termes fournie par ANA. Le taux de rappel approximatif est donc de 35%.

En ce qui concerne la précision, sur les 2 085 candidats termes sortis par ANA nous en avons retenu 961, ce qui équivaut à un taux de précision de 46,09%.

Dans ce qui suit, nous illustrons les critères appliqués dans la validation des candidats termes.

2.2.2. Critères retenus pour la validation des candidats termes

Afin de procéder à la validation des candidats termes, nous avons dû établir des critères pour distinguer les résultats pertinents. Outre la pertinence sémantique, nous avons pris en considération la pertinence syntaxique, c'est-à-dire les limites du découpage en ce qui concerne les termes complexes. Pour le critère de pertinence sémantique, nous avons retenu toutes les séquences ayant un statut terminologique dans le corpus, c'est-à-dire que le choix n'a pas été limité aux techniques et aux moyens thermaux, mais a été élargi également à des termes de la médecine et d'autres domaines connexes au domaine thermal (chimie, pharmacologie)[147]. Lorsqu'une séquence affichait un caractère quelque peu douteux, nous l'avons recherchée dans le corpus à l'aide du menu Locate Pattern du logiciel Unitex, car les résultats de l'extraction par ANA ne comportaient pas de concordances.

En tête de liste, on trouve 83 suites de chiffres et de chiffres et de mots, extraits à tort comme candidats termes, comme par exemple les suites *000 cures*, *10 à 20 minutes*, *2009 thermes*. Comme on peut l'imaginer, ces candidats termes ont tous été rejetés[148]. De même, nous avons rejeté :

[147] Nous précisons que ce critère sémantique a été établi lorsque nous avons annoté à la main la portion de corpus choisie pour le calcul du rappel. Comme l'opération d'annotation a précédé la phase de validation des candidats termes, pour des raisons de cohérence nous l'avons appliqué aux résultats d'ANA.

[148] Nous renvoyons à l'Annexe D pour une liste de ces candidats termes que nous n'avons pas retenu.

- les unités terminologiques complexes incomplètes : *Agence Nationale d'Accréditation et d'Evaluation* à la place de *Agence Nationale d'Accréditation et d'Evaluation de la Santé*, *bain avec eau* à la place de *bain avec eau thermale* ou *bain avec eau courante* ;

- les séquences contenant au moins un terme mais qui résultent d'un mauvais découpage[149] : *b troubles urinaires, conclusion la cure, coxarthrose et la gonarthrose, cure est un moment privilégié*;

- les mots simples qui devraient faire partie d'unités polylexicales non terminologiques : *faveur* à la place de *en faveur* ;
- les mots ou suites de mots introduits par des adjectifs numéraux ou ordinaux : *deux essais, première étude* ;
- les séquences de mots introuvables dans le corpus telles qu'elles ont été repérées par ANA : *piscine bain, bain douche en immersion bain, coût efficacité*;
- les mots ou suites de mots qui appartiennent à la langue générale et qui n'ont pas de sens spécifique dans le domaine d'étude : *point de vue*, *âge moyen* ;
- les mots ou suites de mots étrangers : *balneotherapy hydrotherapy*.

En revanche, les candidats termes retenus appartiennent aux typologies suivantes :

- les termes, simples et complexes, afférents au thermalisme et désignant :
    1) des techniques de soin ou des traitements : *bain, cataplasmes de boue* ;
    2) des moyens thermaux : *boue, eau minérale, algues thermales* ;
    3) des objets : *maillot de bain* ;
    4) des structures : *piscine thermale* ;
- les termes médicaux, simples et complexes, désignant :
    1) des parties du corps : *hanche, appareil digestif, muqueuses buccales* ;
    2) des pathologies : *gonarthrose, incapacité fonctionnelle* ;
    3) des paramètres et des critères utilisés dans les études médicaux : *aveugle du patient, indice algo fonctionnel* ;
    4) des soins et des médicaments : *interventions chirurgicales, prothèse,*
    5) des organismes concernant la santé : *Haute Autorité de Santé* ;
    6) des professions : *médecins thermalistes, kinésithérapeute* ;
    7) des secteurs de la médecine : *hydrologie, chirurgie dentaire* ;
    8) des entités humaines : *population de curistes* ;
    9) des actions : *accompagnement, accueil* ;
- les termes désignant des substances chimiques : *radon, soufre, CO2 naturel* ;
- les préfixes ayant une pertinence dans la formation de termes médicaux comme *gastro* ;
- les termes désignant des structures : *centre de soins, hôpital thermal* ;
- les verbes et syntagmes verbaux à l'infinitif affichant une pertinence avec le domaine d'étude : *prescrire une cure thermale*.

[149] Outre les séquences mal découpées contenant au moins un terme, dans la liste figurent de nombreuses séquences non terminologiques résultant elles aussi d'un mauvais découpage : *aider à mieux, début et la fin.*

2.2.3. Quelques remarques sur l'expérience avec ANA

Cette expérience d'analyse du corpus PTC par le logiciel ANA, indépendamment des résultats obtenus pour ce qui concerne les taux de rappel et de précision, se révèle intéressante pour les quelques points qu'elle nous permet de développer ci-dessous.

En II.1.5.1., nous avons vu que la première opération effectuée par le logiciel sur le corpus lors du traitement est une opération de nettoyage, suite à laquelle le corpus se présente dépourvu de signes diacritiques et signes de ponctuation. L'élimination des signes diacritiques n'est pas un avantage pour le traitement d'une langue comme le français, qui en fait un large usage dans la distinction des homographes. Prenons par exemple deux candidats termes figurant dans la liste : *derives* et *classes en soins*. Le premier peut être tant le substantif féminin pluriel *dérives* que le substantif masculin pluriel (ou aussi l'adjectif masculin pluriel) *dérivés*, alors que l'interprétation du second peut osciller entre *classes en soins* et *classés en soins*. L'effacement des signes de ponctuation n'est pas non plus sans conséquences sur la qualité des candidats extraits, comme le démontre la séquence suivante, extraite comme un seul candidat : *bain aerobain bain douche en immersion bain douche sous marine douche d eau thermale terebenthinee douche forte pression immersion en piscine illutation locale multiple illutation generale compresse etuve locale sudation en cabine individuelle.* En réalité, cette séquence contient une liste de termes de la même catégorie sémantique – des techniques thermales – qui apparaissent sous forme de liste dans le corpus et sont séparés par des virgules. Un autre exemple à l'appui de cet argument est le candidat *faveur de la cure thermale une amelioration*, dont le repérage dans le corpus donne comme résultat : « Ces essais montrent, en faveur de la cure thermale, une amélioration significative ».

Une autre limite mise en évidence par l'extraction est représentée par le fait que le logiciel n'opère pas de distinction entre les séquences lexicales et les séquences numériques : or, ces dernières, bien que souvent très récurrentes dans des corpus textuels, ont rarement un statut terminologique, en tout cas dans la plupart des domaines. Il existe sans aucun doute des termes de certains domaines scientifiques (comme la chimie, les sciences mathématiques ou même la médecine parfois) qui contiennent des chiffres, mais ce n'est pas valable pour tous les corpus ou les domaines. C'est une autre limite liée au fait que le logiciel s'appuie uniquement sur la fréquence et considère les caractères numériques à l'instar des caractères alphabétiques.

La non-utilisation de ressources linguistiques conduit à des résultats de qualité limitée de plusieurs points de vue.

Tout d'abord, comme le logiciel ne reconnaît pas de catégories grammaticales, les candidats termes extraits sont donnés toutes catégories confondues, sans aucune distinction et sans contextes, et comme il ne comporte pas de concordancier, cela oblige le traducteur/terminologue à mener des recherches supplémentaires pour l'attribution des catégories grammaticales, opération nécessaire dans la confection d'un dictionnaire, par exemple.

Outre les erreurs de découpage – souvent grossières –, le logiciel a extrait comme CT bon nombre de séquences qui font partie d'expressions adverbiales figées ou semi-figées, dont l'utilisation est fréquente dans les discours spécialisés.

De plus, nous avons constaté que certaines séquences extraites – qui à première vue pourraient sembler acceptables – n'existent pas comme telles dans le corpus. Si l'on se limitait à la validation des CT de la liste sans vérifier la concordance d'un terme dans le corpus, on risquerait de produire des termes fictifs.

Il y a aussi un dernier point qui mérite d'être développé. Nous avons vu plus haut (figure 1) que pour chaque candidat extrait le logiciel fournit le nombre d'occurrences et les segments de texte d'où le candidat terme a été extrait avec le nombre d'occurrences pour chaque segment. Or, on constate des approximations. Par exemple, pour le candidat *affections neurologiques* – outre les segments *affection neurologique* et *affections neurologiques* – sont donnés aussi les segments *affections nephrologiques* et *affections urologiques* qui, de plus, n'ont pas été extraits comme CT en raison de leur basse fréquence (=1) mais sont pourtant deux autres termes du corpus. L'explication de ce regroupement malencontreux est probablement l'hypothèse selon laquelle *néphrologiques* et *urologiques* résulteraient de fautes de frappe. Autre exemple d'approximation, sur la base des segments suivants : *bains au radon* (2 occurrences), *bains avec radon* (1), *bains de radon* (3) et *bains sans radon* (3), c'est ce dernier le candidat extrait.

Peut-être les performances d'ANA varient-elles considérablement en fonction des corpus et des domaines, s'il est vrai que le traitement de deux corpus de taille inférieure au nôtre (120 000 et 30 000 mots) ayant trait respectivement au nucléaire et à la commercialisation du miel a donné des résultants nettement plus satisfaisants (75% des termes validés par les experts dans le premier cas, 80% dans le deuxième), à en croire ENGUEHARD (1993 : 383-384). Ou bien il se peut que la personne chargée de la validation des candidats termes ait été moins rigoureuse que nous dans le choix des termes à retenir, incluant même des termes mal découpés ou les termes suivis de chiffres. Nous ne pouvons pas le savoir. En revanche, nous croyons pouvoir affirmer que ce type de logiciel est peu adapté si l'on vise des résultats de qualité avec des corpus de taille modérée qui contiennent de nombreux termes dont les occurrences sont peu nombreuses, comme le démontrent le taux élevé de silence (la contrepartie du taux de rappel approximatif, donc 65%) et la qualité des formes extraites dont nous avons parlé plus haut. Surtout pour ce qui concerne ce dernier aspect, le recours à des ressources linguistiques s'avère indispensable.

2.3. Expérience d'extraction terminologique par le logiciel TermoStat

Contrairement à ANA, bien que développé initialement pour un projet d'entreprise, TermoStat est disponible en ligne et il est possible de le tester en créant son propre espace personnel sur le site de référence[150]. Pour la description du logiciel, nous renvoyons à II.1.5.3., où nous avons traité les logiciels basés sur des approches hybrides.

Nous avons soumis les mêmes textes que nous avions proposés pour le traitement par ANA, Alger B_019 et Corpus PTC. Les deux ont pu être analysés.

Il est possible d'exporter les résultats de l'extraction sous forme d'un fichier texte ; en tout cas, à moins que l'utilisateur ne les supprime, les analyses des corpus textuels sont gardées par défaut dans son espace personnel.

[150] http://olst.ling.umontreal.ca/~drouinp/termostat_web/index.php. Le logiciel peut uniquement être testé, il ne peut pas être téléchargé.

Pour le prétraitement du texte, TermoStat recourt à l'étiqueteur morphosyntaxique TreeTagger, qui procède à la désambigüisation des mots susceptibles d'appartenir à plusieurs catégories grammaticales. Suite à la phase de prétraitement, chaque unité du texte se voit assigner une seule étiquette syntaxique. L'étiquetage morphosyntaxique permet à l'utilisateur de mener des recherches plus ciblées. En fait, depuis l'interface de TermoStat on peut choisir la catégorie grammaticale des termes à extraire (nom, verbe, adjectif ou adverbe), outre le type des unités terminologiques (simples, complexes ou les deux).

Nous avons limité la recherche aux unités terminologiques nominales, tant simples que complexes. L'extraction a donné une liste de 175 candidats termes pour le premier texte et 3 522 pour le deuxième. Les critères appliqués pour la validation sont les mêmes que ceux utilisés dans l'expérience avec ANA. Sur les 175 CT extraits pour le texte Alger B_019, nous en avons retenu 95, pour un taux de précision de 54,28%. En ce qui concerne Corpus PTC, 1 769 CT ont été validés sur les 3 522 extraits, pour un taux de précision égal à 50,22%[151]. Il nous semble correct de préciser qu'une bonne partie des CT rejetés ont été exclus parce qu'ils comportaient un déterminant[152]. Ce qui équivaut à dire que si nous avions adopté des critères moins rigoureux, le taux de précision aurait été plus élevé de quelques points.

2.3.1. Présentation des résultats de l'extraction dans Termostat

L'interface des résultats dans TermoStat permet d'accéder à cinq fenêtres différentes : Liste des termes, Nuage, Statistiques, Structuration et Bigrammes. Dans la première, les données sur chaque candidat terme sont organisées en cinq colonnes. Dans la première colonne (candidat de regroupement) sont listés les candidats termes, suivis de leur nombre d'occurrences dans le texte (fréquence) et de leur score de spécificité[153] (colonne score (spécificité)). La quatrième colonne (variantes orthographiques) liste les variantes que le logiciel a repérées pour chaque candidat, alors que la dernière colonne (matrice) en décrit la structure syntaxique. Pour faciliter la compréhension de cette description, nous nous servons de l'image ci-dessous[154] :

[151] Pour l'évaluation de TermoStat, nous prenons en considération uniquement l'évaluation du taux de précision. Nous avons sacrifié le calcul du rappel à l'analyse de la structuration de quelques termes particulièrement significatifs du domaine, ce qui nous semblait plus intéressant.

[152] Il semble que le logiciel ait souvent du mal à extraire des CT commençant par une voyelle sans le déterminant qui les précède.

[153] Il s'agit de mesures statistiques prises en compte dans l'identification des PLS (Pivots Lexicaux Spécialisés), dont nous avons déjà parlé (§II.1.5.3.).

[154] Bien que l'image ne montre que des matrices Nom ou Nom Adjectif, le logiciel reconnaît des structures syntaxiques plus complexes.

Figure 2 : fenêtre Liste des termes.

Les termes sont listés sous forme de liens hypertextuels, qui permettent d'accéder à deux autres fenêtres : une fenêtre de contextes et une autre de concordances (le logiciel affiche les 250 premières occurrences d'une concordance).

La fenêtre Nuage affiche la liste des 100 termes dont le score de spécificité est le plus élevé dans le corpus sous forme de nuage. Le score de spécificité des candidats termes est exprimé par la taille de la police utilisée :

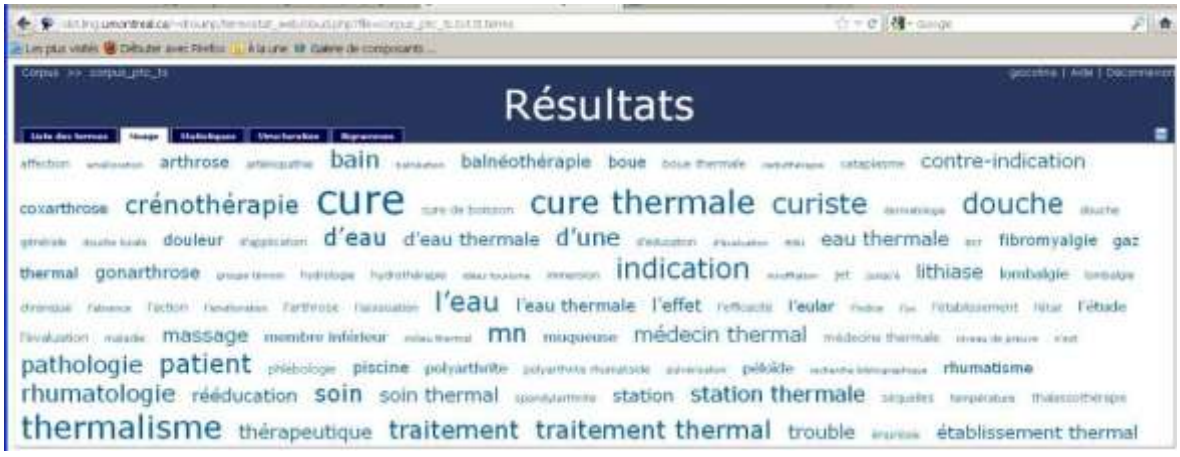

Figure 3 : fenêtre Nuage.

En ce qui concerne la fenêtre Statistiques, elle affiche des données numériques sur le texte analysé, notamment le nombre de candidats termes extraits et la répartition de ces candidats termes suivant les structures syntaxiques (dénommées *matrices* dans le programme), avec nombre exact et pourcentage dans le corpus traité. Le nombre d'occurrences fourni pour chaque matrice contient un lien hypertextuel qui affiche quelques exemples de termes simples particulièrement productifs au niveau des composés.

Les statistiques fournies à propos du texte de petite taille, Alger B_019, sont les suivantes :

Nombre de termes: 175

Matrices
Nom= 74 (42 %)
Nom Adjectif= 54 (31 %)
Nom Préposition Nom= 18 (10 %)
Nom Préposition Nom Adjectif= 9 (5 %)
Nom Nom= 7 (4 %)
Nom PPAdj[155]= 5 (3 %)
Nom Adjectif Adjectif= 4 (2 %)
Nom Préposition Nom Préposition Nom= 1 (1 %)
Nom PPAdj Préposition Nom= 1 (1 %)
Nom Préposition Nom Préposition Nom Adjectif= 1 (1 %)
Nom PPAdj Adjectif= 1 (1 %)

Voici maintenant les statistiques du Corpus PTC :

Nombre de termes: 3522
Matrices
Nom= 1380 (39 %)
Nom Adjectif= 1312 (37 %)
Nom Préposition Nom= 310 (9 %)
Nom Nom= 148 (4 %)
Nom PPAdj= 108 (3 %)
Nom Adjectif Adjectif= 91 (3 %)
Nom Préposition Nom Adjectif= 68 (2 %)
Nom Adjectif PPAdj= 32 (1 %)
Nom Adjectif Préposition Nom= 17 (0.5 %)
Nom Adjectif Préposition Nom Adjectif= 13 (0.4 %)
Nom Préposition Nom Adjectif Adjectif= 9 (0.3 %)
Nom PPAdj Préposition Nom Adjectif= 8 (0.2 %)
Nom Préposition Nom Préposition Nom= 7 (0.2 %)
Nom Adjectif Adjectif Adjectif= 4 (0.1 %)
Nom PPAdj Préposition Nom= 3 (0.1 %)
Nom Adjectif Préposition Nom Adjectif Adjectif= 2 (0.1 %)
Nom PPAdj PPAdj= 2 (0.1 %)
Nom Adjectif PPAdj PPAdj= 1 (0.0 %)
Nom PPAdj Adjectif= 1 (0.0 %)
Nom PPAdj PPAdj PPAdj= 1 (0.0 %)
Nom Adjectif PPAdj PPAdj Adjectif= 1 (0.0 %)
Nom Adjectif Préposition Nom Préposition Nom Préposition Nom= 1 (0.0 %)
Nom Préposition Nom Préposition Nom Adjectif= 1 (0.0 %)
Nom Adjectif PPAdj Adjectif PPAdj= 1 (0.0 %)
Nom Préposition Nom Préposition Nom Préposition Nom= 1 (0.0 %)

[155] Cette abréviation désigne un participe passé à valeur adjectivale, comme *rendu* dans le terme *service médical rendu*.

Passons maintenant à l'analyse de la fenêtre Structuration, qui comporte 3 colonnes. Les deux premières, Candidat de regroupement et Fréquence, sont les mêmes que pour la fenêtre Liste des termes, alors que la troisième, Terme inclus, liste les unités terminologiques complexes que le logiciel a extraites pour chaque candidat de regroupement[156].

Figure 4 : fenêtre Structuration du texte Corpus PTC.

Pour ce module, les performances du logiciel sont directement proportionnelles à la taille du texte soumis à l'analyse. Cela est particulièrement évident si l'on compare la fenêtre « Structuration » des deux textes analysés. En effet, pour le texte de petite taille, TermoStat affiche dans cette liste 25 candidats de regroupement, mais fournit un seul terme rattaché, *soins complémentaires*, sous le candidat *soin* : les 24 autres candidats de regroupement n'affichent aucun terme rattaché, ce qui nous semble résulter d'une anomalie logicielle car parmi les résultats de l'extraction figurent des candidats termes composés qui contiennent certains des 24 autres candidats termes simples, comme par exemple dans les cas de :

- ***source*** : *sources chaudes, source soufrée, stockage de l'eau de source, sources thermales* ;
- ***bain*** : *bains locaux, bains simples, bains carbo-gazeux, bains de vapeur* ;
- ***paraffine*** : *cataplasmes de paraffine, enveloppements de paraffine* ;
- ***eaux*** : *eaux thermominérales* ;
- ***indications*** : *indications thérapeutiques* ;
- ***rééducation*** : *rééducation fonctionnelle, protocoles de rééducation fonctionnelle* ;
- ***affections*** : *affections rhumatismales* ;
- ***centre*** : *centre de thalassothérapie*.

Les résultats sont en revanche bien plus intéressants pour Corpus PTC, comme il est possible de le déduire de la figure 4. Toutefois, une analyse détaillée de ces résultats met encore en évidence des limites. Considérons par exemple la structuration fournie pour

[156] Le candidat de regroupement peut tant être un CT simple qu'un CT complexe.

les termes suivants : *cure, eau, bain, douche*, extraits dans l'ordre que nous donnons, c'est-à-dire par fréquence :

Candidat de regroupement	Terme inclus[157]
cure	semaine de cure – cure de diurèse – orientation de cure – cure de boisson – cure thermale – indication de cure – cure contrôlée – cure hydrominérale – cure de prévention – cure de talassothérapie – fin de cure – cure consécutive – nombre de cures – motif de cure
eau	eau courante – eau chaude – eau minérale – eau thermale – consommation d'eau – eau de mer – eau chlorurée – eaux hyperthermales – eaux mères
bain	bain bouillonnant – bain de gaz sec – bain local – bain de gaz – bain simple – bain d'eau – bain d'eau – bains carbo-gazeux – bains de boue – bain de vapeur – bain complet – bain thermal – bain général – maillot de bain – bain soufrés
douche	douche térébenthinée – douche de vapeur – douche générale – douche sous-marine – douche filiforme – douche locale – douche pénétrante – douche vaginale – douche thermale

Tableau 1 : exemples de l'extraction terminologique complexe fournie par TermoStat pour les termes simples *cure*, *eau*, *bain* et *douche*.

La totalité des termes complexes fournis pour *eau* et *douche* a été validée, alors que 9 termes sur 14 ont été retenus pour *cure* et 13 sur 15 pour *bain*. Dans ce dernier cas, de plus, un même CT apparaît deux fois dans la liste : nous l'avons exclu car il est incomplet. Comme on peut le voir, TermoStat non plus n'est pas exempt d'erreurs de découpage, ce qui est normal, si l'on considère que le découpage des termes est souvent une opération ardue même pour les terminologues.

Plutôt que ce genre d'erreurs, ce qui nous frappe est la quantité réduite de termes complexes fournie pour deux de ces termes simples, dont la fréquence est très élevée dans le Corpus PTC. Nous nous référons à *bain* et *douche*, dont la productivité dans le corpus est extraordinaire[158]. Il suffit de penser que l'analyse des concordances dans le corpus a abouti à 67 termes complexes pour *bain* et 53 pour *douche*. Si le but du terminologue est l'élaboration d'un glossaire des soins thermaux qui soit le plus exhaustif possible, cet

[157] Nous reprenons l'intitulé de colonne fourni dans le logiciel, bien qu'il porte à confusion, car c'est plutôt le candidat de regroupement qui est le terme inclus dans les séquences affichées.

[158] Il suffit de comparer les termes du tableau avec la liste des termes composés relatifs aux techniques et aux moyens thermaux fournie en annexe : la disproportion est frappante. Il est vrai que cette liste a été dressée à partir de la version définitive du corpus (qui compte environ 15 000 mots en plus), mais les textes regroupés dans Corpus PTC sont tous des textes à « haute densité terminologique ». Parmi ces textes il y a aussi le *Guide des bonnes pratiques thermales*, le texte qui contient le nombre le plus élevé de termes de techniques et moyens thermaux. L'écart constaté entre l'extraction terminologique par TermoStat et la liste que nous avons établie à partir des concordances de ces termes dans Unitex est donc un fait significatif, d'autant plus que nous avons répété l'expérience avec la version définitive du corpus et les résultats pour *bain* et *douche* n'ont pas changé.

objectif ne peut pas être poursuivi en se basant uniquement sur les résultats fournis par l'extracteur.

Un autre aspect qui nous a suscité des perplexités a été la fréquence de ces candidats de regroupement reportée dans la fenêtre Structuration : nous avons constaté des écarts – parfois considérables – entre la fréquence affichée par TermoStat et celle identifiée par le concordancier d'Unitex. Nous montrons le nombre d'occurrences repérées par les deux logiciels pour les 10 premiers candidats de regroupement de la liste Structuration dans le tableau ci-dessous :

Candidat de regroupement	Fréquence TermoStat	Fréquence Unitex
cure	1014	1112
traitement	643	647
soin	571	585
patient	449	644
indication	421	591
eau	391	752
thermalisme	340	354
bain	330	607
station	326	326
effet	322	432

Tableau 2 : comparaison entre la fréquence de 10 termes dans TermoStat et Unitex.

Comment expliquer ces écarts ? Dans un premier temps, nous avons fait l'hypothèse que cela pouvait être dû à la flexion. Autrement dit, dans TermoStat les occurrences auraient été calculées soit uniquement sur le singulier soit uniquement sur le pluriel d'un candidat, alors qu'Unitex relevait les deux. Toutefois, des recherches dans le corpus menées par le menu Locate Pattern d'Unitex et limitées d'abord au singulier puis au pluriel de chaque candidat ont démenti cette hypothèse. La même situation s'est vérifiée dans le cas d'une deuxième recherche limitée cette fois-ci aux occurrences de ces candidats commençant par une majuscule.

Passons maintenant à la dernière fenêtre de l'interface du logiciel, Bigrammes. Cette dernière liste les collocations binaires Verbe-Nom que le logiciel a repérées dans l'analyse.

Figure 5 : fenêtre Bigrammes affichée pour le traitement de Corpus PTC.

De même que pour la fenêtre Liste des termes, en cliquant sur les termes affichés dans la fenêtre Bigrammes il est possible d'accéder à une autre fenêtre de Contextes qui affiche les contextes et les concordances. Toutefois, dans quelques cas nous avons constaté que la recherche de contextes et de concordances aboutissait à une fenêtre qui n'affichait aucun résultat, ni pour les contextes ni pour les concordances. Cela s'est vérifié pour les unités commençant par une voyelle et précédées par un déterminant et une apostrophe. De plus, parmi les candidats figurant dans la colonne des verbes il y a des candidats étiquetés à tort comme tels. Sur les 60 bigrammes Verbe-Nom identifiés par le logiciel dans le Corpus PTC, seuls 22 sont corrects.

Un lien hypertextuel en jaune à droite de chaque unité lexicale listée dans la fenêtre Bigrammes permet d'accéder à une autre fenêtre, Décomposition, qui affiche les autres éventuelles collocations dans lesquelles entre l'unité lexicale.

Figure 6 : fenêtre Décomposition.

La figure 6 montre la collocation *constituer* (une) *contre-indication*. Les informations figurant dans la fenêtre Décomposition se lisent de la façon suivante : la ligne Termes en relation affiche l'autre terme de la collocation affiché dans la fenêtre Bigrammes et d'éventuels autres termes de la même catégorie grammaticale (dans le cas de *contre-indication*, le logiciel affiche seulement *constituer*. Mais pour le verbe *jouer*, par exemple, le logiciel affiche comme termes en relation les noms *rôle* et *remise en forme*). L'indice numérique qui suit les termes en relation représente le **score d'association**, c'est-à-dire de la force d'association entre les mots qui composent le bigramme.

2.3.2. Quelques remarques à propos de l'extraction par le logiciel TermoStat

Nous avons vu plus haut que sur les deux textes analysés le taux de précision de TermoStat dépasse 50% et il est susceptible d'augmenter si on inclut parmi les termes validés les CT précédés par un déterminant.

Bien que les résultats de l'extraction par TermoStat ne soient pas parfaits, la qualité des séquences extraites par ce logiciel est nettement meilleure par rapport aux séquences extraites par ANA. Le découpage syntaxique des unités en est la preuve. Nous croyons que cela s'explique par l'utilisation de ressources linguistiques.

Néanmoins, on remarque que, dans ce logiciel aussi, la fréquence constitue un critère fondamental dans l'acquisition des termes. Nous avons eu l'occasion de le constater à plusieurs reprises, comme nous allons le détailler dans ce qui suit.

Tout d'abord, nous avons remarqué que la qualité des séquences extraites semble être directement proportionnelle à leur fréquence : cela ressort avec d'autant plus d'évidence lorsqu'on analyse les CT extraits dont les occurrences sont peu nombreuses (=2), au moins dans le Corpus PTC. Si les CT apparaissant en tête de liste ont de fortes probabilités d'être validés, au fur et à mesure que l'on parcourt la liste les termes corrects sont de moins en moins nombreux. Parmi les CT rejetés figurant en bas de liste il y a des séquences sans aucun caractère terminologique comme *qu'elles*, *s'y adjoint*, *recommandations de l'eular s'est achevée*, pour n'en citer que quelques exemples.

De plus, cet aspect est également ressorti à propos des modules de structuration des deux textes. En ce qui concerne le texte Alger B_019, on se souviendra que le module de structuration n'a presque pas donné de résultats, alors que des candidats termes composés contenant quelques candidats simples de la fenêtre Structuration figuraient parmi les résultats fournis dans la fenêtre Liste de termes. Pour ce qui est du traitement de Corpus PTC, nous avons souligné plus haut que peu de termes composés ont été identifiés à partir de termes simples très productifs dans le corpus sous cet angle, ce qui peut se traduire en un taux de silence élevé.

Cette impression est confirmée lorsqu'on lit sur la page Web qui sert de manuel pour l'utilisateur[159] que la fenêtre Bigrammes « présente les bigrammes les plus ***forts*** du texte analysé »[160]. Nous pouvons affirmer que cela se fait au détriment du rappel, car de nombreuses collocations passent sous silence.

D'autres erreurs affichées par l'extraction sont par exemple ces séquences incorrectes d'un point de vue grammatical (surtout pour des erreurs d'accord de genre et nombre), comme par exemple *climatologie médicales* et *boue local* (qui probablement font partie de séquences plus longues). De même, nous avons constaté des unités d'autres catégories grammaticales extraites à tort comme noms : les adjectifs *sous-marine* et *manosonique* ou l'adverbe *aujourd'hui*. Il arrive également qu'un même candidat terme fasse l'objet d'une double extraction et que les variantes orthographiques d'un même terme ne soient pas regroupées (cela s'explique peut-être par l'étiquetage fourni par TreeTagger).

En général, il nous semble que TermoStat est un logiciel intéressant en vue d'applications terminologiques : c'est un logiciel moderne, dont l'utilisation ne requiert pas

[159] http://olst.ling.umontreal.ca/~drouinp/termostat_web/doc_termostat/doc_termostat.html

[160] Nous soulignons.

de compétences poussées en informatique. De plus, il est possible de mener des recherches ciblées en fonction de la catégorie grammaticale et de repérer les contextes et les concordances d'un candidat – bien que le nombre de ces derniers soit limité à un seuil maximal.

Ce dernier aspect constitue en revanche une limite du logiciel, si les occurrences d'un terme dépassent le seuil de 250, associé au fait qu'il n'est pas possible d'exporter le résultat d'une concordance. Ce qui s'ajoute aux autres limites du logiciel qui sont à attribuer à l'importance accordée au critère de la fréquence, comme nous l'avons déjà illustré plus haut.

Pour cette raison, comme nous l'avions déjà fait lors de la validation des résultats d'ANA, le recours au concordancier d'Unitex s'est souvent révélé nécessaire.

Pour résumer

Dans ce chapitre, nous avons présenté des tests de logiciels d'acquisition de terminologie. Après avoir illustré quelques difficultés liées à ces expériences (§2.1.), notamment en ce qui concerne le repérage ou l'utilisation de ces logiciels, nous avons présenté deux expériences d'acquisition de terminologie à partir de corpus par les logiciels ANA (1993) et TermoStat (2002), dont nous avons fourni une description en II.1.5.1. et II.1.5.3.

Deux textes ont été choisis pour le traitement par les deux outils : un texte de petite taille (environ 1 500 mots) et un corpus d'environ 164 000 mots. Le premier n'a pas pu être traité par le logiciel ANA (outil statistique), alors que ce genre de problème n'a pas été rencontré dans le traitement par TermoStat (outil à base de techniques hybrides). Néanmoins, les performances de ce deuxième logiciel se sont révélées bien meilleures sur le texte de taille supérieure.

Les performances d'ANA (§2.2.) ont été évaluées en termes de rappel et précision (§2.2.1.). Nous avons présenté les critères établis pour la validation des candidats termes (§2.2.2.) et nous montrons des exemples de CT validés et de CT rejetés, en motivant nos choix sur la base des critères de pertinence sémantique et syntaxique appliqués. Comme les résultats de l'extraction ont été peu satisfaisants, nous en avons conclu que le seul critère de la fréquence est insuffisant lorsqu'on vise des résultats de qualité avec des corpus qui présentent de nombreux termes dont la fréquence est peu élevée(§2.2.3.).

L'utilisation combinée de calcul probabiliste et ressources linguistiques, telle qu'elle est faite dans le logiciel TermoStat (§2.3.), a montré des résultats bien meilleurs que le seul recours au calcul probabiliste, surtout en ce qui concerne le découpage des termes. Pour ce deuxième logiciel, qui est accessible au public – à la différence du premier – nous avons fourni également quelques informations sur les différents modules (§2.3.1.).

Les deux expériences ont mis en évidence que les extracteurs sont des outils limités, surtout s'ils sont dépourvus de concordancier, comme dans le cas d'ANA, et si la fréquence est le critère principal qui inspire le logiciel dans l'acquisition des termes.

L'attribution du statut terminologique à une unité lexicale est une opération qui souvent dépend en large partie de l'intuition du traducteur/terminologue et qui donc résiste

à l'automatisation. Pour cette raison, il est nécessaire de disposer d'outils plus souples, qui laissent une plus grande liberté à l'utilisateur (§2.3.2.).