Le logiciel Unitex présentation et possibilités d'exploitation pour le travail terminographique

Le recours à des outils informatiques est désormais incontournable en terminographie, au point que L'HOMME affirme que

> « Actuellement, toute recherche portant sur des termes fait appel à une forme ou une autre de traitement informatique, si bien que la distinction entre ***terminotique*** et ***terminographie*** ne se justifie que dans un contexte pédagogique. » (2004 : 17)

Comme nous l'avons vu dans I.3., depuis le début des années 1990 de nombreux logiciels ont été développés pour automatiser au moins en partie le travail terminographique. Basés sur des méthodes statistiques, symboliques ou hybrides, les logiciels conçus dans une optique terminographique accomplissent généralement les tâches suivantes : extraction de candidats termes, création de réseaux notionnels (ontologies), alignement de termes (pour les travaux sur plusieurs langues).

Dans ce chapitre nous présentons le logiciel Unitex, qui n'a pas été conçu expressément pour la terminologie mais qui peut révéler des possibilités d'exploitation non négligeables dans ce domaine.

1.1. Les besoins des terminologues : des concepts, des termes, des traductions

Pour décrire un domaine, les terminologues ont besoin de repérer les concepts qui le structurent et les termes utilisés pour exprimer ces derniers. Les types d'informations recherchées sur les termes peuvent varier en fonction de l'application visée, mais généralement ces informations sont : la définition du terme, le contexte d'usage, les variantes éventuelles de ce terme. Les définitions sont souvent élaborées par les terminologues sur la base des informations repérées dans le(s) corpus textuel(s) de support à leurs recherches. Le contexte, en revanche, est typiquement une portion – une phrase, un passage – du corpus dans laquelle le terme apparaît, pour en montrer l'usage en discours. Quant aux variantes, ce sont les différentes dénominations qu'un concept peut se voir attribuer. Il est important de les identifier et de les répertorier pour éviter de produire des entrées doubles pour un même concept (que ce soit dans un glossaire ou dans une banque de terminologie). Lorsqu'un terminologue se trouve face à deux ou plusieurs variantes pour un même concept, il opère une sélection sur la base des critères suivants : la fréquence et la répartition dans le corpus, le niveau de langue (la langue standard est préférée aux variétés régionales ou argotiques), la primauté de l'écrit sur l'oral. Comme le courant traductionnel est un des courants les plus productifs en terminologie, les terminologues ont aussi besoin de traductions d'une langue source vers une ou plusieurs langues cibles.

1.2. Unitex : un outil à base de méthodes symboliques

Développé par Sébastien Paumier (2002), Unitex est un logiciel qui réunit différents programmes pour le traitement de textes en langues naturelles sur la base de ressources lexicales. Plus précisément, il s'agit de ressources issues des travaux du lexique-grammaire – dictionnaires électroniques, des tables et des grammaires locales – qui, grâce au réseau RELEX, ont été étendus à d'autres langues. Le logiciel, dont la dernière version est la 3.1 bêta[122], est téléchargeable sous une licence LGPLLR depuis le site du LIGM[123]. Les langues actuellement disponibles dans l'outil sont : l'allemand, l'anglais, l'arabe, le coréen, l'espagnol, le finnois, le français, le géorgien ancien, le grec (ancien et moderne), l'italien, le norvégien, le polonais, le portugais (du Portugal et du Brésil), le serbe, le russe et le thaï. Unitex n'a pas besoin d'un système opérationnel précis pour être utilisé : il marche tant sur Windows que sur Linux et Macintosh OS.

Le logiciel accepte des données brutes (c.-à-d., qui n'ont pas été prétraitées), la seule condition à respecter pour pouvoir analyser un texte est de le coder en Little-Endian Unicode[124]. Il n'y a pas de restrictions sur la taille des textes non plus : étant basé sur des méthodes symboliques, le logiciel peut analyser même des textes courts, à la différence des outils statistiques.

En outre, le logiciel est désormais partie intégrante du projet Gramlab, qui vise à mettre à disposition des entreprises des logiciels libres d'accès et gratuits[125]

Nous renvoyons au manuel d'utilisation[126] pour les procédures d'installation du logiciel. Le manuel, tout comme les menus du logiciel, est en langue anglaise.

1.3. La phase de prétraitement

Lorsque l'on utilise Unitex pour la première fois, le logiciel demande à l'utilisateur de choisir un répertoire de l'ordinateur où il veut stocker ses données. Si l'on travaille sur plusieurs langues – comme dans notre cas – un répertoire différent est créé pour chaque langue. Dans ce répertoire, le logiciel installe six dossiers : Cassys, Corpus, Dela, Elag, Graphs, Inflection. L'utilisateur place dans le dossier Corpus les fichiers à soumettre à l'analyse : il sera ainsi possible de les sélectionner depuis le menu Text.

[122] La version que nous exploitons dans ce travail est la version 3.0.
[123] http://infolingu.univ-mlv.fr/
[124] Il est possible de transcoder les fichiers dans Unitex par le biais de la fonction Transcode files depuis le menu File edition.
[125] Pour plus d'informations, consulter : http://www.gramlab.org/fr/home.html
[126] Téléchargeable depuis le site du LIGM.

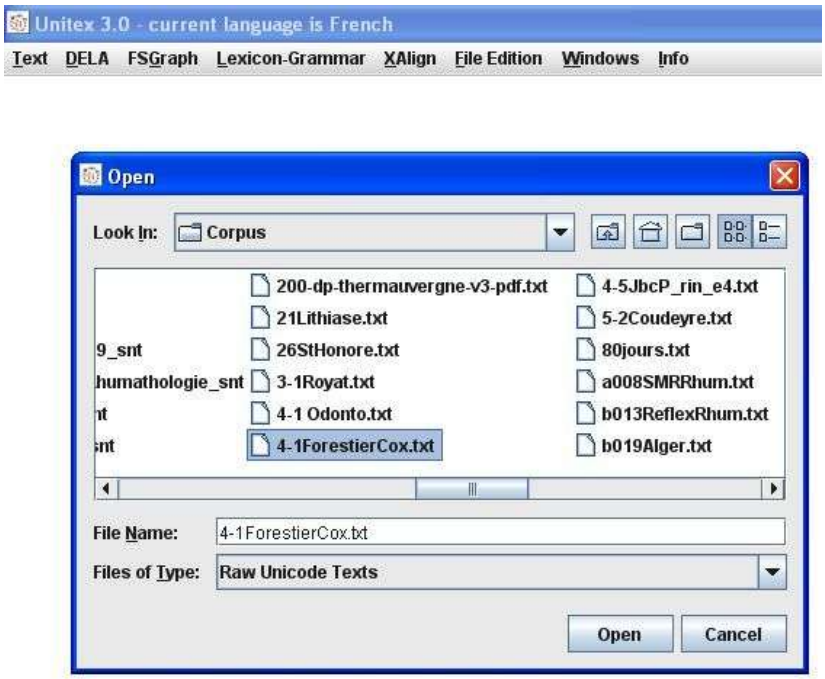


Figure 1 : sélection du texte à traiter

Une fois ouvert le texte, le logiciel ouvre une interface qui demande si l'on veut prétraiter le texte. Par défaut, le logiciel applique au texte les dictionnaires disponibles pour la langue choisie. Si c'est nécessaire à de l'analyse, on peut demander de produire aussi l'automate du texte à la fin de l'étape de prétraitement, en cochant la case Construct Text Automaton (en bas à gauche)[127] :

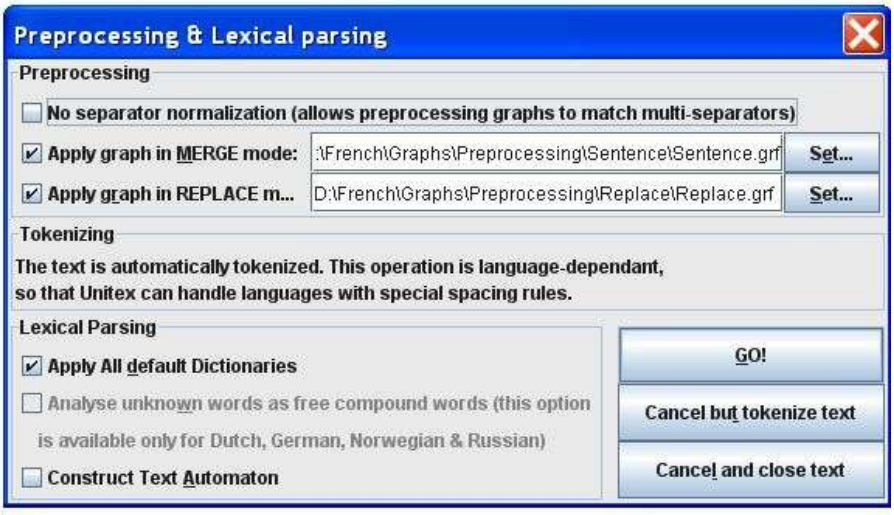


Figure 2 : l'interface permettant l'accès au prétraitement.

Trois opérations fondamentales sont exécutées pendant la phase de prétraitement : le comptage des formes du texte, l'étiquetage de ces formes, la segmentation du texte en phrases. Les résultats de ces opérations sont affichés dans trois fenêtres différentes. Ainsi,

[127] La construction de l'automate du texte peut aussi être demandée aussi par la suite, en choisissant Construct FST-Text depuis le menu Text.

dans la première fenêtre (Token List) sont données toutes les formes[128] présentes dans le texte (signes diacritiques inclus) avec le nombre d'occurrences. Il est possible d'afficher la liste par fréquence (ordre décroissant) ou par ordre alphabétique. La deuxième fenêtre, Word Lists, est divisée en trois sous-fenêtres : une contenant les mots simples, une autre listant les formes composées (dans ces deux premiers cas, il s'agit des formes reconnues par les dictionnaires appliqués) et une dernière dans laquelle sont listées toutes les formes non reconnues par les dictionnaires.

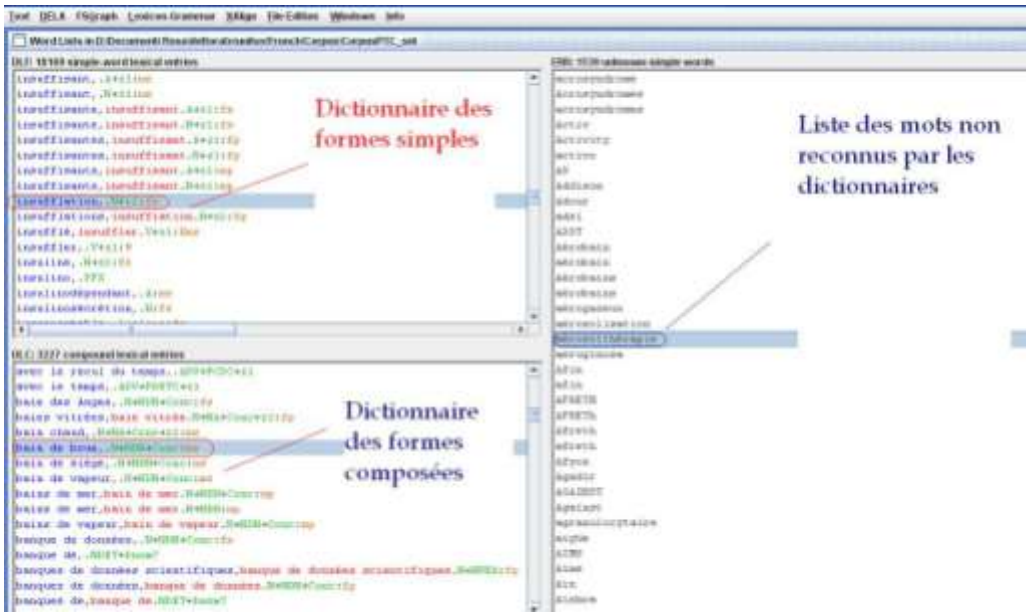


Figure 3 : la fenêtre Word Lists.

Les formes étiquetées se différencient des formes inconnues tout d'abord par l'utilisation de couleurs : bleu, rouge, vert et jaune. Elles sont suivies d'une série de codes morphosyntaxiques : outre la catégorie grammaticale, pour chaque forme sont donnés, dans le cas du dictionnaire DELA du français, le genre et le nombre (personne, mode et temps dans le cas des verbes). Certaines formes se voient attribuer une étiquette sémantique (par exemple, *Conc* pour les noms concrets). Des informations sur la fréquence du mot dans la langue sont explicitées par l'étiquette « z », qui prend des valeurs numériques de 1 à 3, selon que le mot est un mot très courant (z1), un mot spécialisé (z2) ou un mot très spécialisé (z3). L'entrée est en bleu, la forme canonique en rouge, les informations sur la catégorie grammaticale, la fréquence et d'éventuelles étiquettes sémantiques sont en vert, alors que les informations relatives à la flexion sont en jaune. En guise d'exemple, considérons les deux entrées suivantes dans le DELA du français :

1) buccale,buccal.A+z2:fs ;

2) sels minéraux,sel minéral.N+NA+Conc+z1:mp.

L'exemple 1) nous décrit la forme *buccale* : c'est un adjectif féminin singulier, dont la forme canonique est *buccal* et l'emploi relève d'un niveau de langue quelque peu technique (z2). L'exemple 2) définit une forme composée, *sels minéraux* : il s'agit d'un

[128] *Token* ne correspond pas à *mot* : il s'agit de n'importe quel caractère du texte, même un espace est un *token*.

nom composé masculin pluriel (la forme canonique est *sel minéral*) du type Nom-Adjectif, qui désigne un nom concret et dont l'emploi est fréquent dans la langue.

Suite aux dernières modifications opérées sur le logiciel, il est désormais possible de vérifier rapidement si une unité lexicale fait partie de la nomenclature des dictionnaires électroniques. Depuis le menu DELA, on sélectionne Lookup : on accède ainsi à une fenêtre qui permet de choisir les dictionnaires dans lesquels on veut vérifier une unité lexicale donnée :

Figure 4 : fenêtre Dictionary Lookup.

En ce qui concerne la liste des mots inconnus, elle peut être très intéressante dans une optique terminographique : souvent, bon nombre des formes inconnues sont des néologismes ou des termes propres à un domaine donné, qui ne sont donc pas inclus dans les dictionnaires électroniques. La liste peut se révéler un point de départ pour le repérage des termes que l'on recherche. D'autres formes inconnues peuvent être des entités nommées[129] – qui présentent elles aussi un intérêt du point de vue terminographique –, des fautes de frappe et des mots étrangers.

La troisième et dernière fenêtre présente le texte segmenté en phrases : le symbole {S} (= *sentence*) délimite une portion de texte que le logiciel a reconnu comme une phrase, sur la base des signes diacritiques (notamment, les points) et des lettres capitales.

Une fois achevée la phase de prétraitement, on peut commencer à mener des recherches en exploitant les informations fournies en fonction de l'application visée.

1.4. À la recherche des termes, simples et composés

La priorité des terminologues est la recherche des concepts et des termes utilisés pour les décrire, comme nous l'avons vu au début de ce chapitre.

[129] Les entités nommées sont les noms propres comme des noms de personne ou d'institutions, des toponymes ou encore des marques.

Pendant la phase de documentation, les terminologues établissent « l'arbre du domaine », qui sert à hiérarchiser les concepts en génériques, spécifiques et associés (qui au niveau linguistique sont des hyperonymes, des hyponymes et des co-hyponymes). Ils recherchent ensuite les termes désignant ces concepts dans les corpus textuels. Il y a différentes façons de mener la recherche des termes : si l'on recourt à un extracteur de terminologie, on peut procéder à la validation des candidats-termes sur la base des occurrences dans le corpus. Si le but du travail est la mise à jour d'une terminologie, on recherche les termes contenus dans des index ou d'autres produits terminographiques dans des corpus textuels plus récents, de façon à éliminer les termes périmés et ajouter les éventuels néologismes.

Les termes se répartissent en simples et composés, selon qu'ils sont formés d'un seul ou de plusieurs mots au sens typographique (les termes composés par soudure comme *balnéothérapie* sont des termes simples). D'habitude, les termes-clés d'un domaine sont ceux qui interviennent dans la formation des composés. Ces derniers, dont le nombre dépasse de loin les termes simples, demandent une attention particulière de la part des terminographes pour plusieurs raisons. Tout d'abord, le découpage du terme : il faut bien établir les limites du terme ; cela peut être d'autant plus compliqué dans le cas des termes surcomposés, qui sont des composés contenant un ou plusieurs termes composés. Ensuite, il faut considérer que les structures syntaxiques de ces composés peuvent varier selon les domaines : si les structures du type *Nom Adjectif* ou *Nom de Nom* (au moins pour la langue française) sont très productives pour tous les domaines, l'identification des structures des surcomposés nécessite d'une analyse plus approfondie. L'étude des structures syntaxiques des composés est d'autant plus importante dans une optique contrastive : non seulement ces structures peuvent changer dans le passage d'une langue source à une langue cible (songeons par exemple que les structures *Nom Adjectif* du français sont plutôt des structures *Adjectif Nom* en anglais), mais il arrive qu'un terme soit un composé dans la langue source et un terme simple dans la langue cible et vice-versa (c'est le cas de *maladie cœliaque* en français, traduit en italien par *celiachia*).

Dans les paragraphes suivants, nous montrons comment nous avons procédé à la recherche des termes – simples et composés – à partir des résultats du prétraitement des corpus dans Unitex.

1.5. Un programme capital : Locate Pattern

Depuis le menu Text on accède à Locate Pattern, un programme qui sert de "moteur de recherche" sur le texte analysé. C'est une fonctionnalité capitale pour les possibilités qu'elle offre au travail terminographique en raison de l'extrême souplesse qu'elle permet à l'utilisateur dans la recherche. Le programme Locate Pattern est, en outre, étroitement lié à l'application de grammaires locales, comme nous le verrons lorsque nous présenterons celles-ci.

Pour l'instant, nous nous limitons à illustrer la recherche d'expressions régulières. Si l'on veut localiser une ou plusieurs formes du texte étiquetées par le prétraitement, il faut utiliser Locate Pattern. La recherche peut être menée de façons différentes. On peut rechercher, par exemple :

- une unité lexicale donnée, comme *bain*. Si l'unité fait partie du dictionnaire électronique, l'utilisation des chevrons (<bain>) permet de repérer tant la forme canonique que les formes fléchies. La recherche d'une unité sans chevrons donne dans les résultats uniquement la forme recherchée. L'omission des chevrons est d'ailleurs obligatoire lorsqu'on veut localiser une forme listée dans la fenêtre des mots inconnus ;

- toutes les occurrences d'une catégorie grammaticale : <A> pour tous les adjectifs ;

- un masque lexical : <douche.N> recherche toutes les occurrences du nom *douche* dans le texte (et non pas douche comme forme fléchie du verbe *doucher*)[130]

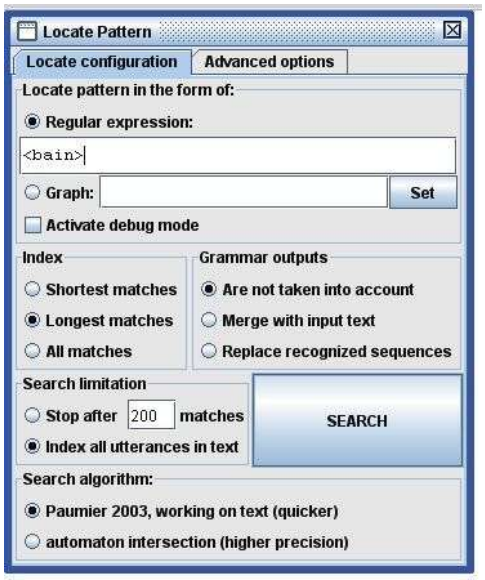
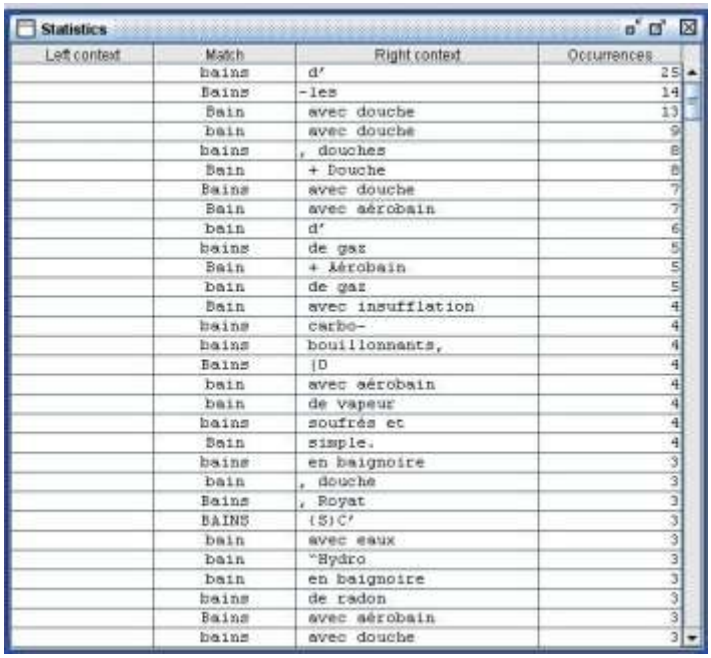


Figure 5 : la recherche de toutes les occurrences du mot *bain* (au singulier et au pluriel) dans la fenêtre Locate Pattern.

Comme on peut le voir dans la figure 5, il est possible d'afficher toutes les occurrences du mot dans le texte ou de limiter la recherche à un nombre restreint (par défaut, le logiciel se limite à un seuil de 200 formes). Une fois choisis les paramètres de la recherche, on lance celle-ci en cliquant sur Search. Une autre fenêtre s'affiche, celle des séquences localisées (Located sequences). Depuis cette dernière fenêtre, on choisit d'afficher soit une concordance du mot dans le texte soit des statistiques à propos des contextes du mot recherché. Il suffit de délimiter le nombre des tokens à gauche ou à droite du mot recherché.

[130] Ce type de recherche est possible à condition d'appliquer un système de levée d'ambigüités.



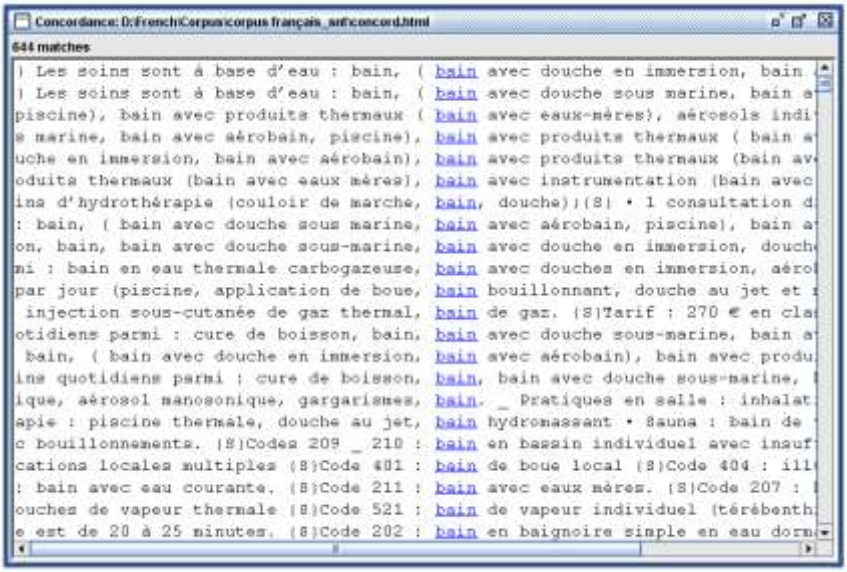
Statistics

Left context	Match	Right context	Occurrences
	bains	d'	25
	Bains	-les	14
	Bain	avec douche	13
	bain	avec douche	9
	bains	, douches	8
	Bain	+ Douche	8
	Bains	avec douche	7
	Bain	avec aérobain	7
	bain	d'	6
	bains	de gaz	5
	Bain	+ Aérobain	5
	bain	de gaz	5
	Bain	avec insufflation	4
	bains	carbo-	4
	bains	bouillonnants,	4
	Bains	{D	4
	bain	avec aérobain	4
	bain	de vapeur	4
	bains	soufrés et	4
	Bain	simple.	4
	bains	en baignoire	3
	bain	, douche	3
	Bains	, Royat	3
	BAINS	{S}C'	3
	bain	avec eaux	3
	bain	"Hydro	3
	bain	en baignoire	3
	bains	de radon	3
	Bains	avec aérobain	3
	bains	avec douche	3

Figure 6 : statistiques des contextes du mot *bain*.

Dans la figure 6, nous montrons les statistiques par fréquence du mot *bain* : les valeurs sélectionnées pour la définition des contextes sont de 0 pour le contexte gauche et de 2 tokens pour le contexte droit. L'affichage des statistiques peut constituer le point de départ d'une piste pour le repérage des termes composés.

Si par contre on choisit d'afficher une concordance du mot dans le texte, le résultat se présente de la façon suivante :



Concordance: D:\FrenchCorpus\corpus français_ant\concord.html

644 matches

) Les soins sont à base d'eau : bain, (bain avec douche en immersion, bain
) Les soins sont à base d'eau : bain, (bain avec douche sous marine, bain a
piscine), bain avec produits thermaux (bain avec eaux-mères), aérosols indi
e marine, bain avec aérobain, piscine), bain avec produits thermaux (bain a
uche en immersion, bain avec aérobain), bain avec produits thermaux (bain av
oduits thermaux (bain avec eaux mères), bain avec instrumentation (bain avec
ins d'hydrothérapie (couloir de marche, bain, douche){S} • 1 consultation d
: bain, (bain avec douche sous marine, bain avec aérobain, piscine), bain a
on, bain, bain avec douche sous-marine, bain avec douche en immersion, douch
mi : bain en eau thermale carbogazeuse, bain avec douches en immersion, aéro
par jour (piscine, application de boue, bain bouillonnant, douche au jet et
injection sous-cutanée de gaz thermal, bain de gaz. {S}Tarif : 270 € en cla
otidiens parmi : cure de boisson, bain, bain avec douche sous-marine, bain a
bain, (bain avec douche en immersion, bain avec aérobain), bain avec produ
ins quotidiens parmi : cure de boisson, bain, bain avec douche sous-marine,
ique, aérosol manosonique, gargarismes, bain. _ Pratiques en salle : inhalat
apie : piscine thermale, douche au jet, bain hydromassant • Sauna : bain de
c bouillonnements. {S}Codes 209 _ 210 : bain en bassin individuel avec insuf
cations locales multiples {S}Code 401 : bain de boue local {S}Code 404 : ill
: bain avec eau courante. {S}Code 211 : bain avec eaux mères. {S}Code 207 :
ouches de vapeur thermale {S}Code 521 : bain de vapeur individuel (térébenth
e est de 20 à 25 minutes. {S}Code 202 : bain en baignoire simple en eau dorm

Figure 7 : affichage d'une concordance du mot *bain* dans le texte.

Pour chaque concordance recherchée, un fichier HTML nommé "concord" est produit dans le dossier .snt du texte traité et il est consultable indépendamment d'Unitex. Les occurrences s'affichent sous formes de liens hypertextuels, comme on peut le voir dans la figure 7. On accède au contexte d'utilisation d'une forme en cliquant sur le lien de référence. Le mot ou la séquence recherché apparaîtra surligné en bleu (figure 8). Comme

le champ "contexte" est un des plus importants dans les fiches terminologiques, on comprend toute l'utilité de cette opération.

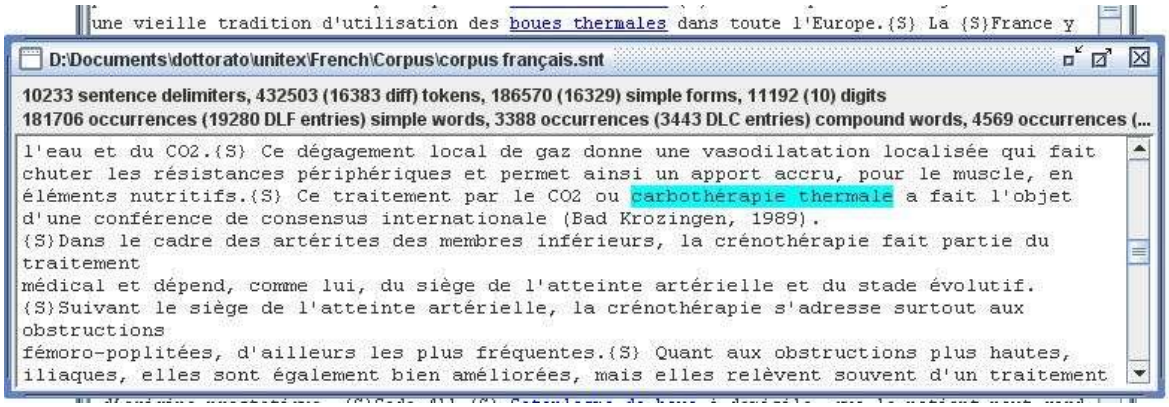


Figure 8 : affichage du contexte de la séquence *carbothérapie thermale*.

L'affichage des concordances reste invariable pour tous les types de motifs recherchés. D'autres types de recherches, définies comme des recherches par des symboles spéciaux ou *métas* (PAUMIER 2011 : 72), sont également possibles. Toutefois, nous croyons qu'encore un autre type de recherche apporte davantage d'intérêt dans une optique terminologique : il s'agit de la recherche par application de filtres morphologiques. Parmi tous les types de filtres morphologiques illustrés dans le manuel, deux ont retenu notre attention :

1) <<^>> : ce filtre permet de repérer toutes les formes commençant par une suite donnée de caractères : par exemple, <<^hydr>> reconnaît toutes les formes débutant par *hydr* ;

2) <<$>> : ce filtre, en revanche, permet de reconnaître toutes les formes ayant une même portion finale. Ainsi, <<thérapie$>> retrouve toutes les séquences du texte se terminant par *thérapie*.

Les deux sont utiles dans l'étude d'un champ sémantique déterminé ou des termes composés par soudure mais qui sont à considérer comme des termes simples.

En dernier lieu, la recherche par patrons syntaxiques peut également être lancée depuis Locate Pattern par expressions régulières. Cependant, l'emploi de patrons syntaxiques révèle un potentiel majeur dans des grammaires locales, qui permettent des recherches mieux ciblées.

Si l'on veut comparer les concordances obtenues par deux recherches différentes, il suffit de cliquer sur Show difference with previous concordance – et non pas sur Build concordance – lors de la deuxième recherche. Les différences entre les deux recherches s'affichent dans une nouvelle fenêtre, Concordance Diff, et sont rendues visibles par le biais de différentes couleurs : en violet, une même occurrence apparaissant dans les deux concordances ; en rouge, une occurrence apparaissant de façon plus longue dans l'une des deux concordances ; en vert, une occurrence apparaissant dans une seule des deux concordances. Ci-dessous, nous montrons une fenêtre Concordance Diff résultant de l'application de deux graphes similaires mais non identiques :

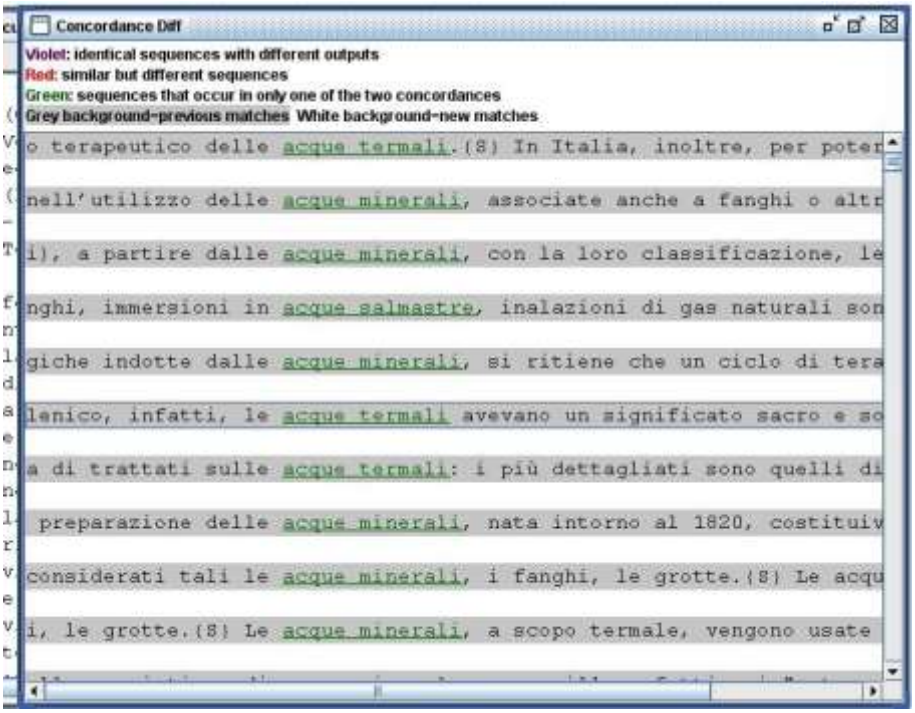


Figure 9 : fenêtre Concordance Diff affichant les différences entre deux concordances consécutives.

1.6. Identification des structures syntaxiques des termes composés

Dans le domaine de la médecine thermale, nous avons identifié deux sous-domaines sémantiques très riches en termes : les moyens thermaux (qui sont des entités concrètes comme des substances) et les techniques thermales (des soins et des méthodes). À l'intérieur de ces deux sous-domaines nous avons identifié quelques termes très fréquents et très productifs au niveau des composés, que nous pouvons considérer comme des hyperonymes. Ensuite, nous les avons utilisés pour identifier des structures syntaxiques des composés dans les deux langues d'étude. Les termes à haute fréquence dans le corpus français sont : *cure* (1331 occurrences), *eau* (832), *bain* (644), *douche* (405). Pour le corpus italien, l'identification des structures des composés a été menée à partir des termes suivants : *acqua* (1176 occurrences), *cura* (365), *fango* (202) et *bagno* (121). Nous avons également retenu l'adjectif *thermal* (2143) et son équivalent *termale* (760), en raison de leur fréquence élevée et de leur incontestable pertinence au domaine. Dans un deuxième temps, des recherches supplémentaires de structures syntaxiques contenant d'autres noms[131] ont été menées sur des textes de taille plus réduite, afin de fournir une description qui soit la plus exhaustive possible.

Sur le modèle de M. GROSS (1985), nous avons établi des catégories de composés sur la base des *N* contenus dans chaque structure. Comme on peut l'imaginer, les structures à base de 2 *N* ou plus sont pour la plupart des structures de surcomposés. La langue française semble être plus productive dans ce sens que la langue italienne, au moins pour ce qui concerne les composés désignant les méthodes et les moyens thermaux : 30 classes

[131] Noms désignant toujours des méthodes ou des moyens thermaux.

syntaxiques de noms composés ont été identifiées contre 17 classes pour l'italien. De son côté, l'italien semble utiliser davantage d'adjectifs dans les termes composés.

1.6.1. Catégories des termes composés du corpus français

Les 30 classes syntaxiques identifiées dans le corpus français se répartissent de la façon suivante : 5 classes syntaxiques à un seul *N*, 13 classes contenant 2 *N*, 11 classes contenant 3 *N* et une seule classe à 4 *N*. Nous fournissons un exemple pour chaque structure ci-dessous :

Structures à un seul *N*

N A: bain bouillonnant
N A A : douche nasale gazeuse
N PFX-A : douche sous-marine
N PFX-A A : douche sous-marine carbogazeuse
N PFX-A PFX-A : douche sous-marine carbo-gazeuse

Structures à 2 N

N A A Prép N : bain carbogazeux général en baignoire
N A PrépDét N : douche générale au jet
N A Prép N : bain hydromassant en piscine
N PFX-A Prép N A : injection sous-cutanée de gaz thermal
N Prép Dét N : massage sous l'eau
N Prép N A A : bain en eau thermale carbogazeuse
N Prép N ADV A : bain d'eau naturellement bicarbonatée
N Prép N PFX-A : bain avec douche sous-marine, douche de gaz loco-régionale
N Prép N : bain avec aérobain
N Prép PFX- N : bain avec hydro-massages
N PrépDét N : douche au jet
N N[132] *: douche affusion*
N A N : douche forte pression

Structures à 3 N :

N A PrépDét N Prép N A : douche locale au jet d'eau thermale
N A Prép N A Prép N : bain local de gaz sec de Royat
N Prép A N Prép N : douche de forte pression sous immersion
N Prép N A Vpp Prép N : bain de limon thermal suivi de douche
N Prép N A Prép N A : bain en baignoire simple en eau dormante
N Prép N A Prép N : bain en eau salée avec hydromassage
N Prép N Prép Dét N A : bain de siège pour les pathologies hémorroïdaires
N Prép N Prép N A: bain de vapeur d'eau thermale
N Prép N N[133] *: bain avec eaux mères*

[132] La structure *NN* connaît également la variante *N-N : douche-massage*. Il est important de tenir en compte ce type de variantes lors de l'édition des graphes dans Unitex.
[133] Une autre variante de la structure est *N Prép N-N : bain avec eaux-mères*.

N Prép N Prép N : bain avec insufflation de gaz

Structure à 4 *N* :

N Prép N Prép A N Prép N : douche sous immersion à haute pression en piscine

1.6.2. Catégories des termes composés du corpus italien

Pour ce qui concerne les classes syntaxiques des termes composés du corpus italien, nous en avons identifié 5 à un seul *N*, 9 à 2 *N* et 3 à 3 *N*. Ce faible nombre s'explique aussi par l'existence de quelques variantes orthographiques : songeons par exemple à la classe *N A A A*, qui présente les deux variantes *N A A-A* et *N A-A-A*. Nous avons décidé de garder une structure unique mais de tenir compte des variantes lors de l'édition des graphes. Parfois il a été nécessaire de modifier la description d'une structure d'après la codification des unités dans le DELAF : nous nous référons surtout aux adjectifs composés du type *cloruro-sodico*. Plutôt que créer de nouvelles entrées (dans ce cas, le préfixe *cloruro*), nous avons préféré adapter la description de la classe aux entrées existantes (pour ce même cas, le N *cloruro*). Voici les classes repérées dans le corpus :

Structures à un seul N

N A : fanghi clorurati
N A A[134] *: fango sorgivo sulfureo*
N A A A[135] *: ciclo irrigatorio vaginale termale*
PFX-N: foto-balneoterapia
N PFX-A: cure idro-diuretiche

Structures à 2 N

N A Prép N A : terapie inalatorie di acque termali
N A A Prép N A[136] *: terapia inalatoria termale con acqua sulfurea*
N Prép N: doccia a soffione
N Prép N A : terapia con acqua sulfurea
N Prép N A A : idromassaggio con acqua termale salsobromoiodica
N Prép N A A A : bagno con acqua minerale clorurosodica solfata
N PrépDét N A: irrigazioni del cavo orale
N-N: balneo-crenochinesiterapia
N A N-A A: acqua minerale cloruro-sodica sulfurea

Structures à 3 N

N N Prép N: irrigazione goccia a goccia
N A N-A-N-A A: acqua minerale solfato-sodica-calcio-magnesiaca termale
N Prép N A-N-A : inalazione di acqua sulfureo-solfato-calcica

[134] Cette structure a aussi la variante *N A-A: inalazione caldo-umida*.

[135] Deux variantes ont été repérées pour cette structure : 1) *N A A-A: antroterapia naturale caldo-secca* et 2) *N A-A-A: acqua alcalino-sodico-calcica*.

[136] La variante *N A-A Prép N A : inalazione caldo-umida con acqua solfurea* est également attestée.

1.7. Les graphes : réalisation et possibilités d'exploitation

Lors de notre introduction au lexique-grammaire, nous avons parlé des grammaires locales, qui sont des transducteurs représentables sous forme de graphes d'automates finis, pour reprendre la définition de M. GROSS (1997). Il est fondamental de savoir réaliser et manipuler les graphes que l'on peut créer pour les nombreuses possibilités qu'ils offrent à l'utilisateur. Des dossiers différents sont prévus pour sauvegarder les graphes en fonction de leurs possibilités d'exploitation. Avant d'illustrer ces possibilités, nous donnons quelques indications sur la création des graphes.

1.7.1. Graphes syntaxiques

Pour créer un graphe, il faut sélectionner New dans le menu FSGraph. Il s'affiche une fenêtre où apparaissent une flèche sur la gauche et un point sur la droite : la flèche indique le début du chemin du graphe, le point est en revanche la fin de ce chemin. Les différentes étapes du chemin d'un graphe sont représentées par des boîtes, que l'on crée par Ctrl+clic. Lorsqu'une boîte est créée, il faut lui attribuer un contenu : tout comme pour les expressions régulières illustrées pour le programme Locate Pattern, ce contenu peut être un mot, une catégorie grammaticale, un masque lexical, un filtre morphologique, etc. Une boîte peut contenir plusieurs unités à la fois : il suffit de les séparer par des "+", comme nous le montrons dans la figure 9, qui montre un graphe reconnaissant toutes les séquences contenant les termes *étuve*, *irrigation* et *massage* suivis d'un adjectif. Nous rappelons que pour la recherche par catégories et codes grammaticaux il faut toujours s'en tenir au codage des unités dans le dictionnaire utilisé, sachant que les codes utilisés peuvent changer d'une langue à l'autre[137]. Pour appliquer un graphe au texte, il suffit de cliquer sur Graph – et non pas sur Regular expression – dans la fenêtre Locate Pattern et de choisir le graphe, qui aura été sauvegardé de préférence dans le dossier Graphs.

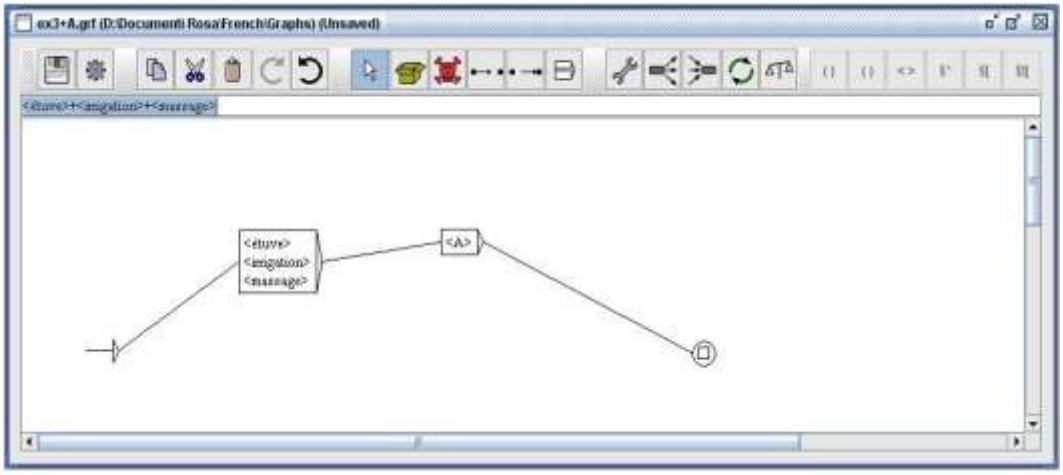


Figure 10 : exemple de graphe à deux boîtes.

Un graphe peut appeler un autre graphe, qui dans ce cas sera appelé sous-graphe. Pour faire appel à un sous-graphe, on crée une boîte dans laquelle le caractère ":" précède

[137] Juste pour citer un exemple, les préfixes sont codés PFX dans le DELA du français et PX dans le DELA de l'italien.

le nom du graphe. On pourra reconnaître le sous-graphe dans la boîte car il sera indiqué en gris, comme le montre la figure 11 :

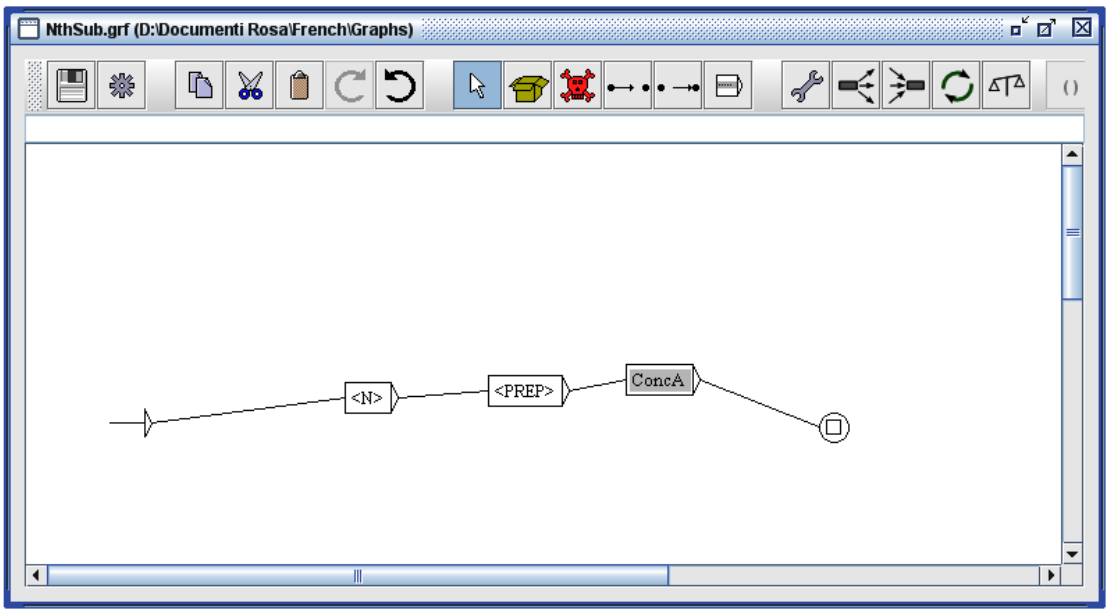


Figure 11 : graphe appelant le sous-graphe ConcA.

L'appel à des sous-graphes peut être utile dans la recherche de termes surcomposés : le graphe de la figure 11, par exemple, reconnaît les séquences nominales suivies d'une préposition et d'un nom composé (nom concret + adjectif).

L'homonymie est à la base de nombreux cas d'ambigüités : une même forme peut appartenir à plusieurs catégories grammaticales et être codée dans les DELA pour chacune de ces catégories (songeons à *douches*, pluriel du substantif féminin *douche* et 2ème personne singulière du présent de l'indicatif du verbe *doucher*). Cela peut produire du bruit dans les résultats des recherches. Toutefois, il y a la possibilité de peaufiner les grammaires locales sans devoir recourir à la désambigüisation sur l'automate du texte (application des grammaires ELAG), dont la manipulation pourrait ne pas être immédiate pour des utilisateurs peu expérimentés. La délimitation de contextes dans le graphe est un moyen de peaufiner le graphe et par conséquent de mieux cibler les recherches. On peut définir des contextes tant positifs que négatifs. Pour délimiter un contexte droit positif, il suffit de cliquer sur l'icône "$[" dans la barre des icônes. Si l'on veut en revanche délimiter un contexte droit négatif, il faut cliquer sur l'icône "$![". La seule condition nécessaire est que son début et sa fin apparaissent dans le même graphe. La figure 12 reproduit un graphe pour lequel deux contextes droits négatifs ont été définis : la recherche vise le repérage des séquences constituées d'un nom et d'un adjectif dans le texte. Le nom doit être une forme simple et non pas une forme composée, ni constituer le début d'un mot composé déjà répertorié dans le dictionnaire : les formes étiquetées comme N dans le dictionnaire des formes composées (CDIC) du type *station thermale* ne seront donc pas reconnues. Pour ce qui concerne le deuxième contexte négatif, le graphe doit identifier les adjectifs du texte

qui ne sont ni l'adjectif *est*[138], ni des participes passés, ni des adjectifs composés, et qui ne constituent pas le début d'un mot composé. Il est également possible de définir des contextes gauches positifs (avec une autre convention), mais pas de contextes gauches négatifs.

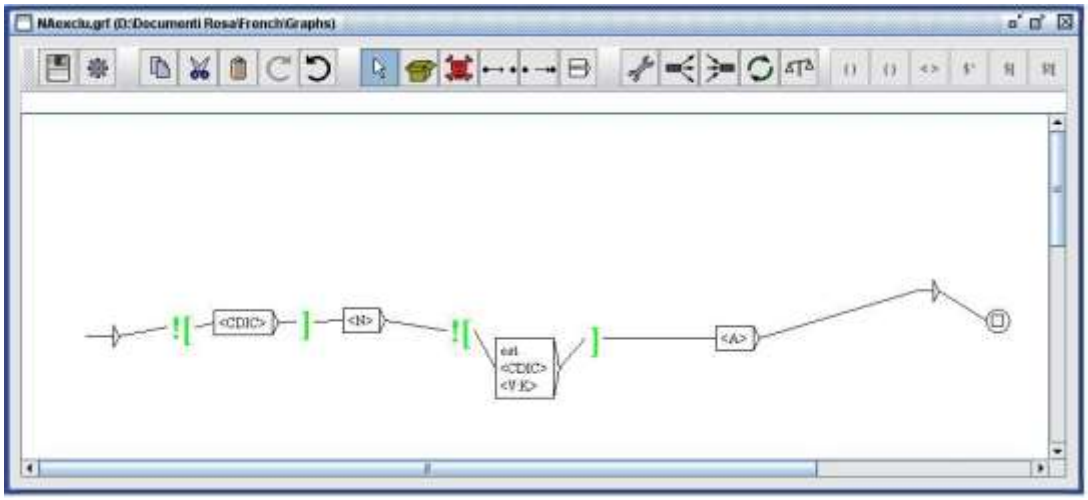


Figure 12 : graphe comportant deux contextes droits négatifs.

Comme on peut le voir, les graphes que nous avons présentés jusqu'ici sont des graphes syntaxiques très simples, dont la réalisation ne requiert pas de compétences très poussées en informatique. Au fur et à mesure que l'utilisateur acquiert de la familiarité avec le logiciel, il sera capable de produire des graphes syntaxiques de plus en plus élaborés. De toute façon, ce qu'il ne faut pas perdre de vue dans la réalisation des graphes est l'attention à la syntaxe des requêtes. En cas de doute, on peut toujours recourir au manuel. Dans le cas de discours spécialisés se rapprochant du stéréotype – comme les bulletins boursiers ou des textes juridiques comme les règlements – on peut formaliser des grammaires locales capables de reconnaître des vastes portions de corpus (voir par exemple NAKAMURA 2005).

1.7.2. Graphes de flexion

Jusqu'à présent, nous avons illustré les graphes servant de grammaires locales, applicables sur un texte depuis la fenêtre Locate Pattern. Nous allons maintenant présenter les graphes de flexion, dont la maîtrise est nécessaire pour la création d'un dictionnaire électronique. Pour pouvoir être utilisés, ces graphes doivent être sauvegardés dans le dossier Inflection et commencer par le code de la catégorie grammaticale pour laquelle ils ont été conçus. Ainsi, les noms des graphes de flexion pour les adjectifs commenceront par *A*, ceux pour les noms par *N* et ceux pour les verbes par *V*. Le code identifiant la catégorie grammaticale est suivi d'un code numérique choisi par l'utilisateur. Voici un exemple de graphe de flexion pour une catégorie d'adjectifs italiens, ceux se terminant en –co/-go (dont *antroterapico* est un exemple de notre corpus) au masculin singulier :

[138] De cette façon, on élimine toutes les occurrences de *est* comme 3ème personne singulière du présent indicatif du verbe *être*, ce qui réduit considérablement le bruit.

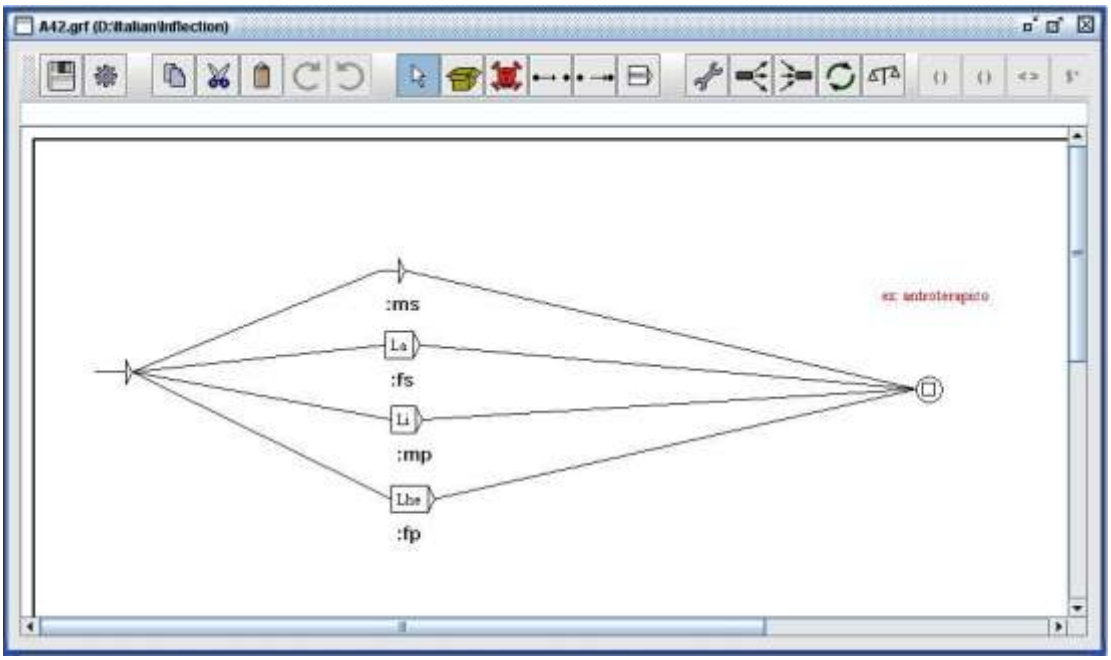


Figure 13 : graphe de flexion A42.

La forme canonique est représentée par une boîte vide, tandis que les formes fléchies sont décrites dans des boîtes[139]. Pour pouvoir assigner une sortie à une boîte, il suffit d'utiliser le caractère "/". Ainsi, pour créer la boîte de la forme canonique masculin singulier, on écrira sur la ligne d'édition : <E>/ms. Le "L" indique que la forme fléchie comporte une lettre en moins à gauche (Left) par rapport à la forme canonique (chaque L indique un caractère : ainsi, si une forme fléchie comporte deux lettres en moins par rapporte à la forme canonique, il faut écrire LL) et il est suivi de la désinence pour une flexion donnée. Les quatre formes produites par le graphe A42 seront ainsi : *antroterapico* au masculin singulier, *antroterapica* au féminin singulier, *antroterapici* au masculin pluriel et *antroterapiche* au féminin pluriel. Pour être opérationnel, un graphe de flexion doit être compilé. Pour cela, il suffit de cliquer sur la deuxième icône en haut à gauche ou d'ouvrir le menu FSGraph>Tools>Compile FST2.

1.7.3. Graphes dictionnaires

Il y a deux façons de générer des entrées de dictionnaire dans Unitex : par l'élaboration de dictionnaires électroniques ou bien par l'édition de graphes dictionnaires. Ces derniers sont particulièrement utiles dans le codage de mots inconnus qui appartiennent à une même catégorie sémantique, comme par exemple des toponymes. Pour générer des entrées de dictionnaire, il suffit de produire une sortie contenant l'étiquette que l'on veut attribuer à la catégorie reconnue par le graphe précédée de la suite ",." (exemple : ,.NPr permet de coder des noms propres). Une fois le graphe terminé, il faut le compiler et le sauvegarder dans le dossier Dela. Voici un exemple de graphe dictionnaire que nous avons élaboré pour le corpus italien :

[139] Si une forme fléchie est identique à la forme canonique (comme dans le cas des adjectifs épicènes), la boîte de la forme fléchie sera également vide.

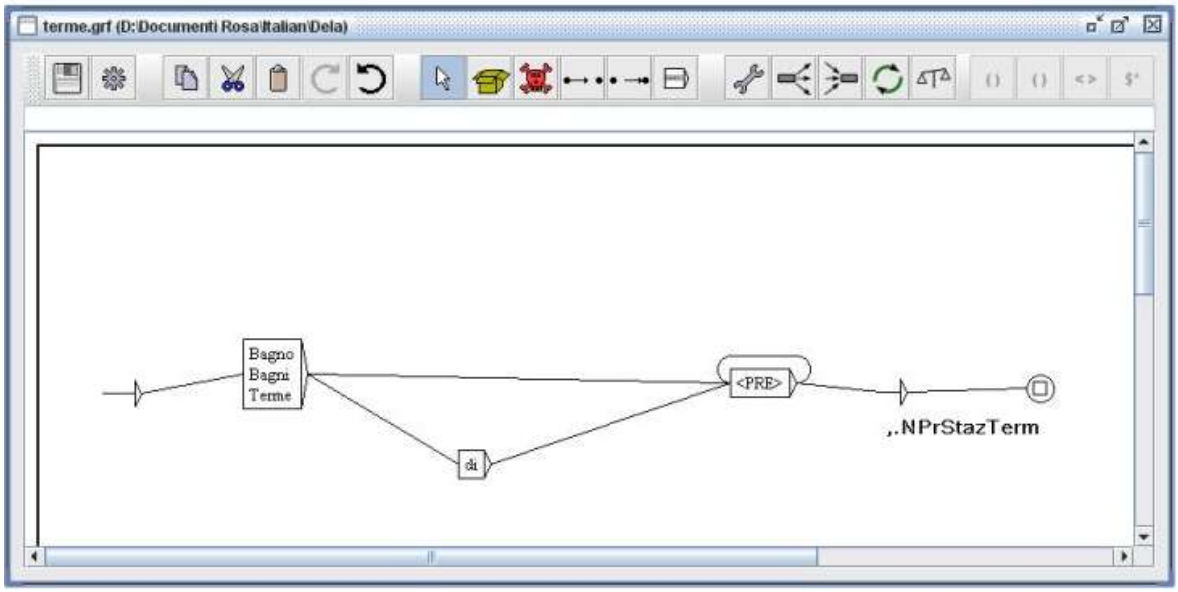


Figure 14 : le graphe dictionnaire terme.fst2 .

Ce graphe reconnaît des noms propres de stations thermales (l'étiquette StazTerm est une abréviation pour « Stazione Termale »), qui sont des formes composées dont le premier élément est *Bagno*, *Bagni* ou *Terme*, suivi d'un nom commençant par une majuscule (<PRE>), qui peut à l'occasion être précédé par la préposition *di*. Nous avons choisi de ne pas mettre entre chevrons les mots *bagno* et *terme* pour pouvoir appliquer ce graphe indépendamment des autres dictionnaires. De plus, la boîte <PRE> comporte une boucle pour permettre au graphe de reconnaître aussi les suites de plusieurs noms commençant par une majuscule.

Pour vérifier les résultats d'un graphe dictionnaire, on l'applique au texte comme un dictionnaire quelconque : il apparaîtra dans la liste des ressources lexicales de l'utilisateur, où on pourra le sélectionner (Text>Apply lexical resources>User resources). Les résultats de l'application du graphe terme.fst2 apparaissent dans la fenêtre des formes composées dans Word Lists :

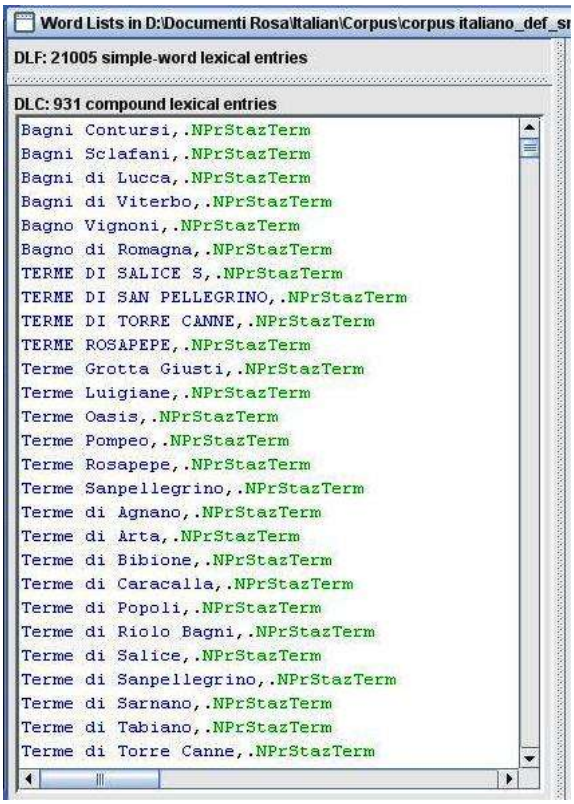


Figure 15 : les formes composées générées par l'application du graphe de la figure 12.

1.8. Création d'un dictionnaire électronique

Malgré l'exhaustivité des dictionnaires DELA, tant du français que de l'italien, après le prétraitement des corpus pas mal de termes propres au domaine ont été listés dans la fenêtre des mots inconnus. Pour qu'Unitex les reconnaisse, nous avons procédé à la création de deux dictionnaires électroniques de mots simples, l'un pour le français et l'autre pour l'italien. Outre les mots non reconnus, nous avons également inclus dans les dictionnaires des termes propres au domaine qui faisaient déjà partie de la nomenclature des DELA, en ajoutant des étiquettes sémantiques très simples pour pouvoir mener des recherches bien ciblées et réduire ainsi le bruit qu'aurait produit une recherche moins raffinée : l'étiquette « Th » pour « thermalisme », « Conc » pour « concret ».

La création d'un dictionnaire électronique dans Unitex se fait depuis le menu File Edition. En cliquant sur New File, on obtient une fenêtre qui permet l'édition de textes[140]. Pour que le dictionnaire soit lisible par Unitex, il faut suivre le codage syntaxique utilisé par le logiciel. Chaque entrée du dictionnaire que l'on veut créer doit faire l'objet d'une ligne. Sur chaque ligne, le terme à inclure doit être suivi d'une virgule, de la catégorie grammaticale d'appartenance et du code numérique qui fait appel à la grammaire de flexion pour ce type d'unité (par exemple, la grammaire de flexion N3 du français permet de fléchir les substantifs masculins comme *bureau*). Il n'est pas nécessaire que les entrées

[140] Cette même fenêtre aussi peut être utilisée pour l'édition de fichiers en format .txt que l'on veut soumettre à l'analyse.

soient listées par ordre alphabétique, car ce sera fait automatiquement par le logiciel une fois que le dictionnaire sera fléchi.

Pour créer nos deux dictionnaires (dicoTher_simple pour le français et dicoTher_simple_ita pour l'italien), nous avons procédé de la façon suivante : tout d'abord, nous avons inclus les termes de la fenêtre des mots inconnus qui semblaient appartenir au domaine. Ensuite, nous avons parcouru la Token List par ordre alphabétique et créé les entrées au fur et à mesure, profitant de la possibilité d'afficher les deux fenêtres côte à côte et en reconstituant le lemme. Pour le choix des unités à inclure, nous avons retenu une fréquence minimale de 3 occurrences, qui est le seuil minimal normalement choisi par les terminologues. En ce qui concerne la pertinence des unités au domaine, le choix a été opéré sur la base des connaissances du domaine acquises pendant la phase de constitution des corpus, qui a servi d'étape de documentation. En cas de doutes nous nous sommes basée sur un ou plusieurs contextes de l'unité « douteuse » dans le corpus, recourant à la fonctionnalité Locate Pattern. De même, des dictionnaires de langue ont été consultés pour valider ou infirmer un choix. Parfois, nous nous sommes trouvée face à des cas problématiques, tel l'adjectif *manosonique*, attesté seulement dans le corpus, sans qu'une définition ne soit disponible. Nous avons décidé de retenir ce terme (et les autres « cas problématiques »), considérant qu'il est plus facile d'éliminer ultérieurement du bruit éventuel plutôt que de compenser le fait d'avoir passé un candidat terme sous silence.

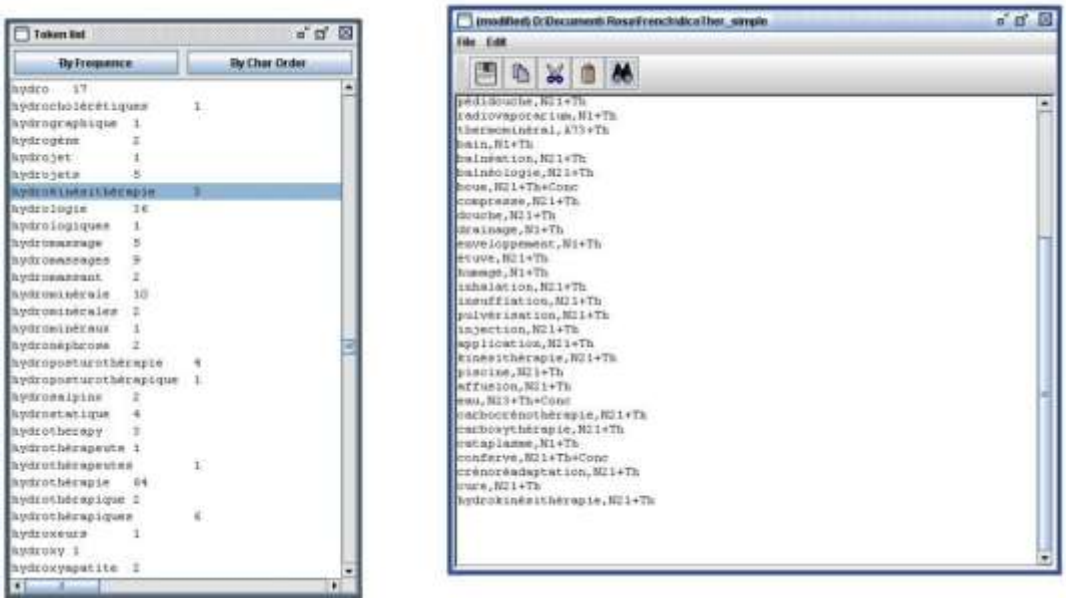


Figure 17 : création du dictionnaire dicoTher_simple.

Une fois le dictionnaire complété, nous l'avons sauvegardé dans le dossier Dela, pour pouvoir ensuite l'ouvrir depuis le menu DELA dans Unitex (DELA>Open), étape fondamentale pour procéder à la flexion du dictionnaire. Après avoir ouvert le dictionnaire, nous l'avons soumis à la procédure de flexion en cliquant sur Inflection.

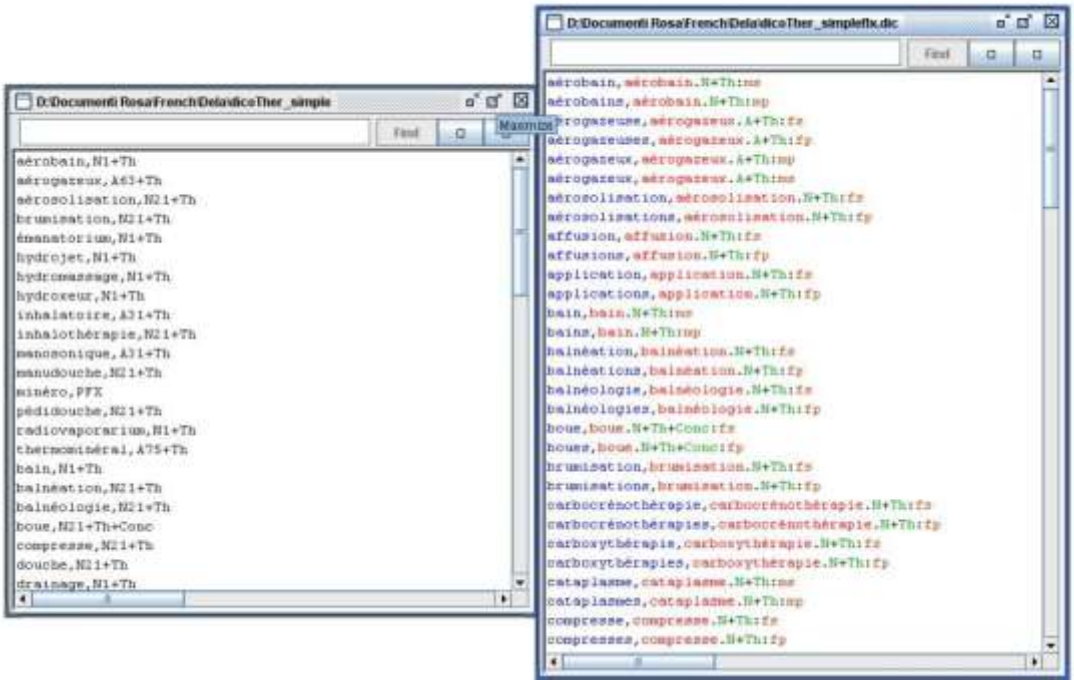


Figure 18 : le dictionnaire dicoTher_simple avant et après la flexion.

Si le dictionnaire n'affiche pas des unités ordonnées alphabétiquement, il suffit de cliquer sur Sort Dictionary dans le menu DELA et les unités seront agencées alphabétiquement. La dernière étape à franchir est la compression du dictionnaire (DELA>Compress into FST). Quand le dictionnaire est fléchi et compressé, on peut l'appliquer au texte en le sélectionnant dans la fenêtre Apply lexical resources depuis le menu Text[141], où il sera listé automatiquement parmi les ressources de l'utilisateur (User resources) en tant que fichier .bin .

On peut établir une priorité dans les dictionnaires à appliquer : cela se fait par l'insertion des symboles "-" et "+" dans le nom du fichier du dictionnaire, avant l'extension .bin. Le symbole "-" indique une priorité élevée, tandis que le symbole "+" indique une priorité faible. L'établissement des priorités des dictionnaires s'avère un moyen pour favoriser la désambigüisation des homonymes : en fait, si une même forme fait l'objet de plusieurs entrées dans des dictionnaires différents, en établissant un dictionnaire prioritaire l'entrée recevra uniquement l'étiquetage prévu par le dictionnaire prioritaire. Voyons cela à l'aide d'un exemple pratique. Plus haut nous avons cité quelques cas d'ambigüité à propos de quelques termes de nos corpus, tels que *cure* et *douche*, étiquetés tant comme *N* que comme *V* dans le DELAF. Ces termes sont décrits en revanche uniquement comme *N* dans notre dictionnaire dicoTher_simple : qualifiant ce dernier de prioritaire, l'étiquetage de *cure* et *douche* – et aussi de *cures* et *douches* – comme verbes fourni dans les autres dictionnaires ne sera plus pris en considération par le logiciel. Ainsi, dicoTher_simple-.bin aura une priorité élevée d'application tandis que dicoTher_simple+.bin aura une priorité faible : ce n'est que dans le premier cas que *cure* et *douche* seront étiquetés uniquement comme *N*.

[141] C'est le même chemin décrit pour l'application des graphes dictionnaires.

1.9. Alignement de textes : le programme X-Align

Unitex est une collection de programmes : parmi ceux-ci il y a aussi le programme X-Align, développé au LORIA en 2006. X-Align est un aligneur de textes : il met en correspondance deux textes, dont l'un est le texte source et l'autre sa traduction dans une autre langue. Il est également possible de travailler sur plus de deux textes à la fois. Chaque texte est découpé en segments, qui peuvent être des phrases, et chaque segment est associé à un ou plusieurs autres. L'alignement se fait donc au niveau de la phrase et non pas au niveau des mots.

Travaillant sur deux corpus comparables, nous n'avons pas pu exploiter ce programme dans le cadre de cette thèse, mais nous croyons utile de donner quelques informations sur X-Align, qui peut se révéler un outil précieux dans le cadre de travaux sur des corpus parallèles, tels que les textes juridiques des communautés bi- ou plurilingues.

Pour accéder au programme, on clique sur le menu X-Align dans Unitex. Le programme demande à l'utilisateur de sélectionner les textes à aligner : ces derniers peuvent être au format .txt – tout comme les textes du répertoire Corpus – ou des textes codés au format TEI (qui est un format XML). Dans la même fenêtre, l'utilisateur doit sélectionner aussi un fichier d'alignement, qu'il aura préalablement construit. Si ce n'est pas le cas, Unitex peut construire une version basique au format TEI d'un fichier texte. Des versions XML des textes sont ensuite fournies sous forme de listes de phrases. Pour les aligner, il faut cliquer sur le bouton Align. A titre d'exemple, nous montrons l'exploitation du programme sur un texte littéraire. Le résultat se présente de la façon suivante :

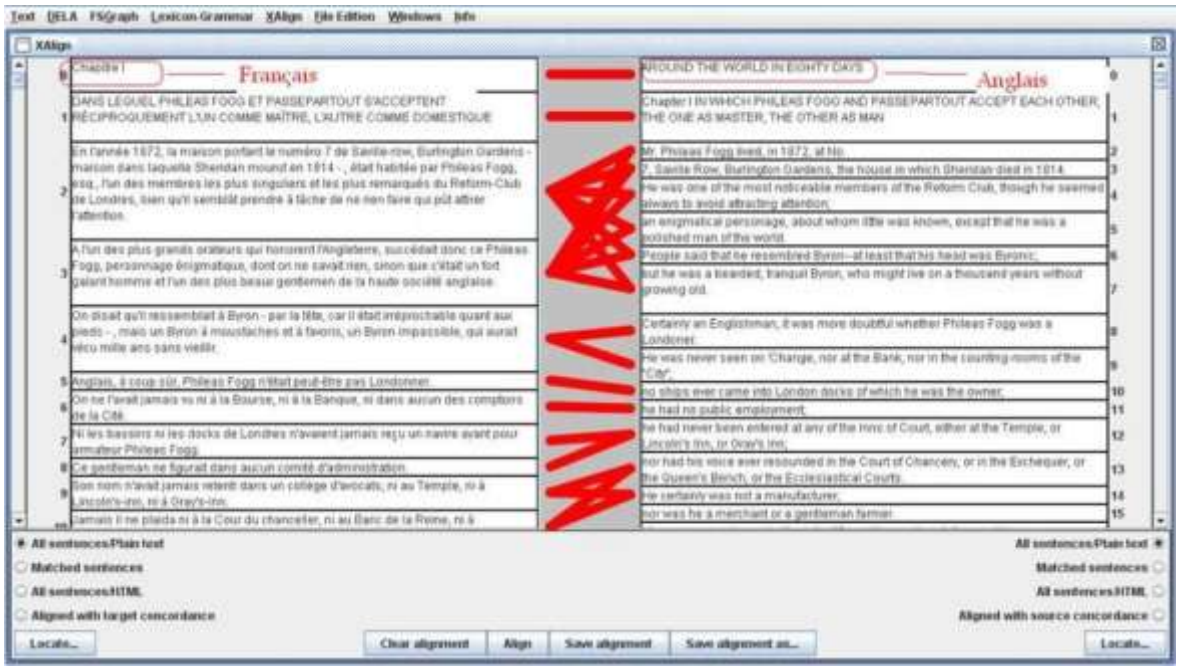


Figure 19 : alignement d'un extrait d'un roman en français et en anglais.

1.10. Exploitation des tables de lexique-grammaire

Nous allons consacrer quelques mots à une autre fonctionnalité d'Unitex que nous n'avons pas exploitée dans ce travail, mais qui pourrait constituer une piste intéressante à

développer. Depuis le menu Lexicon-Grammar on a la possibilité de convertir les tables de lexique-grammaire en graphes. Il s'agit de graphes paramétrés, qui affichent des variables correspondant aux colonnes d'une table de lexique-grammaire. Pour plus d'informations sur le menu Lexicon-Grammar, nous renvoyons au chapitre 9 du manuel.

1.11. Quelques données à propos du travail sur les deux corpus d'étude

Jusqu'à présent, nous avons illustré les possibilités d'exploitation du logiciel Unitex pour la terminologie, parsemant cette présentation de quelques exemples de notre travail sur les deux corpus d'étude. Dans cette section, nous approfondirons quelques aspects liés à la recherche des termes et fournirons aussi quelques données pour ainsi dire numériques.

Un des grands avantages – sans doute le plus grand avantage – des méthodes symboliques est la finesse des descriptions qu'elles produisent, étant basées sur des ressources lexicales et non pas sur des algorithmes. Le taux de silence des recherches par ce type d'outils est donc moins élevé par rapport à des recherches par des outils statistiques. Le revers de la médaille est une augmentation du taux de bruit, c'est-à-dire des résultats non pertinents. Unitex ne fait pas exception dans ce sens. Toutefois, l'extrême souplesse de l'outil permet d'atteindre des résultats plus performants en appliquant quelques méthodes plus ou moins "astucieuses", consistant à identifier des éléments à la base du bruit et mettre en place des stratégies pour les isoler.

Une première cause du bruit est à attribuer au fait qu'Unitex n'opère pas de distinction entre majuscules et minuscules, à moins de ne forcer la casse par le signe « " », ce qui en revanche pourrait produire du silence. Mais voyons cela à l'aide d'un exemple. Dans les deux corpus nous avons rencontré ce problème à propos du même terme : *bain* et son équivalent italien *bagno*. Comme nous l'avons vu plus haut, il s'agit d'un terme très productif en composés. La recherche des concordances pour ce terme a donné parmi les résultats des noms propres de stations thermales françaises et italiennes : *Lamalou-les-Bains, Bains St. Pierre, Bagni Contursi, Bagno di Romagna*, pour n'en citer que quelques exemples. Si le fait de forcer la casse sur les minuscules, d'un côté, excluait ces noms propres, d'un autre il faisait passer sous silence de véritables termes composés apparaissant en début de phrase. En effet, nous avons constaté que souvent un terme se trouve en début de phrase – et comporte ainsi une majuscule – pour faire l'objet d'une définition[142]. La solution pour laquelle nous avons opté a été la création de deux graphes dictionnaires reconnaissant les noms propres de stations thermales, l'un pour le français et l'autre pour l'italien (pour ce dernier, voir figure 14). Cela nous a permis de créer une nouvelle étiquette sémantique pour pouvoir ensuite l'exploiter en tant que contexte négatif dans l'édition de graphes syntaxiques recherchant les termes composés contenant *bain* et *bagno*.

Grâce à une autre étiquette sémantique, Th, nous avons pu limiter la recherche des composés à un nombre restreint de termes. Nous avons introduit cette étiquette lors de la création de nos deux dictionnaires électroniques de termes simples (§1.8.) et nous l'avons utilisée dans les graphes syntaxiques élaborés pour le repérage des termes composés. Nous en montrons l'utilité à partir de quelques données numériques. Soit la structure syntaxique *N A* : sa productivité dans les deux langues analysées ne fait pas de doutes. La recherche de

[142] Le corpus français est particulièrement riche en exemples de ce type.

ce motif sans aucun filtre dans les corpus d'étude a donné comme résultat 20 739 occurrences pour le français et 23 371 pour l'italien. Comme on peut l'imaginer, bon nombre de ces occurrences sont à éliminer, car toutes les séquences nom-adjectif des textes ont été retrouvées, et toutes ne sont pas propres au domaine thermal. Si nous considérons en outre qu'une partie de ces séquences contiennent des mots décrits dans des entrées dédoublées pour homonymie – des tokens étiquetés comme noms ou adjectifs alors qu'ils n'ont pas cette valeur dans les corpus – on comprend qu'il est plus rentable (en termes de temps et de qualité des résultats) d'élaborer des filtres pour peaufiner les recherches plutôt que de valider à la main des listes aussi longues. L'utilisation de l'étiquette Th dans des graphes syntaxiques recherchant les séquences *N A* (où *N* est accompagné de Th : <N+Th>) a réduit les concordances à 1 930 pour le corpus français et à 1 681 pour le corpus italien. Il s'agit de chiffres bien plus raisonnables à contrôler, qui ont pu être réduits davantage par la délimitation de contextes négatifs après une analyse rapide des séquences à éliminer. Nous avons appliqué cette méthode à tous les graphes syntaxiques que nous avons élaborés pour la recherche des composés. Ainsi, pour la recherche des formes composées comportant une ou plusieurs prépositions, nous avons vérifié si toutes les prépositions étaient productives, pour exclure celles qui ne l'étaient pas. Bien évidemment, pour chacune des deux langues nous avons élaboré des filtres adaptés aux types d'erreurs.

Les listes des termes composés français et italiens désignant des techniques et des moyens thermaux sont fournies en annexe par ordre alphabétique, ainsi que des exemples de graphes dictionnaires et de graphes syntaxiques.

Pour résumer

Dans ce chapitre, nous avons introduit le logiciel Unitex, une collection de programmes qui permet de travailler sur de nombreuses langues en exploitant des ressources lexicales telles que des dictionnaires électroniques, des tables de lexique-grammaire et des grammaires locales. Bien qu'il n'ait pas été conçu expressément pour la terminologie, Unitex présente un potentiel d'exploitation intéressant pour satisfaire les besoins des terminologues (notamment, la recherche de termes, contextes et traductions) (§1.1.).

Après avoir fourni quelques données génériques à propos du logiciel (§1.2.), nous en avons illustré le fonctionnement à partir de l'expérience sur nos deux corpus comparables en guise de parcours par étapes. La première étape a été le chargement d'un texte et la phase de prétraitement (§1.3.), qui produit trois opérations fondamentales : le comptage des formes du texte, l'application de dictionnaires électroniques, le découpage du texte en phrases. A cette occasion, nous avons également souligné l'intérêt que peut avoir la fenêtre des mots non reconnus par les dictionnaires électroniques en tant que piste pour le repérage de termes propres à un domaine (§1.4.).

La deuxième étape a été l'introduction au programme Locate Pattern (§1.5.), que nous avons qualifié de « capital », étant donné sa fonction de "moteur de recherche" dans le texte et la diversité des requêtes qu'il permet de mener. Parmi ces dernières, nous avons mentionné les recherches par unité lexicale, par masque lexical, par catégorie grammaticale et par symboles spéciaux (ou *métas*) : en ce qui concerne ceux-ci, une

attention particulière a été accordée à la recherche par filtres morphologiques. C'est à ce moment que nous avons parlé de l'affichage des concordances, des statistiques à propos des contextes d'une unité et de la comparaison entre deux concordances.

A partir de quelques termes très fréquents dans les deux corpus, nous avons identifié les structures syntaxiques des termes composés pour les deux langues de référence, que nous avons regroupées par le nombre de *N* qu'elles contiennent (§1.6., 1.6.1. et 1.6.2.). Cette étape d'identification nous a servi pour introduire les graphes, des transducteurs d'automates à états finis particulièrement performants dans la recherche par motifs syntaxiques. Dans Unitex, les graphes peuvent assumer des fonctions différentes (§1.7.). Trois types de graphes ont été décrits : les graphes syntaxiques (§1.7.1.), les graphes de flexion (§1.7.2.) et les graphes dictionnaires (§1.7.3.). La maîtrise des deuxièmes est indispensable dans la création d'un dictionnaire électronique, qui a été l'étape successive à la présentation des graphes. De cette étape (§1.8.), nous avons montré les critères à partir desquels nous avons procédé à l'édition de deux dictionnaires électroniques de termes simples de la médecine thermale, un pour chaque langue. La création d'un dictionnaire électronique passe par trois jalons : l'édition, la flexion et la compression, qui en permettent la successive application aux textes.

Outre les fonctionnalités d'Unitex que nous avons exploitées dans notre travail sur les corpus, nous avons également consacré quelques lignes au programme X-Align (§1.9.) et au menu Lexicon-Grammarn (§1.10.). Le premier est un aligneur de textes, utile dans les travaux sur corpus parallèles. Le second permet de convertir des tables de lexique-grammaire en une série de graphes paramétrés.

La dernière section du chapitre a fourni quelques données à propos du travail sur les deux corpus d'études, notamment à propos des stratégies mises en place pour réduire le bruit que souvent produisent les recherches par outils linguistiques (§1.11.).