Institut National des Langues et Civilisations Orientales

Département Textes, Informatique, Multilinguisme

Acquisition de connaissances à des fins d'analyse automatique : Extraction des différentes façons de nommer les ingrédients et actifs cosmétiques dans les conversations spontanées des internautes en français et anglais et comparaison multilingue

MASTER

TRAITEMENT AUTOMATIQUE DES LANGUES

Parcours :

Traductique et Gestion de l'Information

par

Karolina KRYGIER

Directeur de mémoire :

Mathieu Valette

Encadrant :

Marguerite Leenhardt

Année universitaire 2017/2018

TABLE DES MATIÈRES

Liste des figures **5**

Liste des tableaux **6**

Résumé **7**

Remerciements **9**

Introduction **11**

I Contexte général **15**

1 État de l'art **17**
1.1 Introduction 17
1.2 Définition et principe 17
1.3 Différentes méthodes 18

2 Méthodes **23**
2.1 Introduction 23
2.2 Etape 1 : Etiquetage morphosyntaxique 24
2.3 Etape 2 : Calcul des distances entre les mots 25
2.4 Etape 3 : Recherche des patrons morphosyntaxiques 26
2.5 Etape 4 : Implémentation des règles des patrons morphosyntaxiques . . 27
2.6 Etape 5 : Nettoyage des expressions relevées pour livrer les candidats termes 28

II Expérimentations **29**

3 Corpus **31**
3.1 Corpus et préparation 31

4 Résultats **33**
4.1 Variantes des éléments de base 33
4.2 Patrons morphosyntaxiques 35
4.3 Candidats termes 37
4.4 Comparaison multilingue des résultats 40

Conclusion générale **43**

Bibliographie **45**

A Annexe : Scripts **47**
A.1 Premier script 47
A.2 Deuxième script 48
A.3 Troisième script 49
A.4 Quatrième script 50

B Annexe : Listes des candidats termes extraits **53**
B.1 **Candidats termes finaux en français** 53
B.2 **Candidats termes finaux en anglais** 66

Liste des figures

0.1 Résultat d'une requête SQL demandant d'afficher les contextes gauche et droit du motif "rose" . . . 11
0.2 Exemple d'expression raccourcie d'un ingrédient . . . 11
0.3 Résultat d'une requête SQL pour afficher les messages comportant le motif "aloe vera" correctement orthographié . . . 12
0.4 Résultat d'une requête SQL pour afficher les messages comportant le motif "aloe" 12
0.5 Extrait de ressources linguistiques - amande douce . . . 12
0.6 Extrait de ressources linguistiques - beurre . . . 13
0.7 Extrait de ressources linguistiques - AHA . . . 13
0.8 Extrait de ressources linguistiques - huile . . . 13
2.1 Liste des éléments de base en français . . . 23
2.2 Liste des éléments de base en anglais . . . 23
2.3 Étiquetage d'éléments de base en français . . . 24
2.4 Étiquetage d'éléments de base en anglais . . . 24
2.5 Calcul de la distance de Levenshtein entre "argan" et "argane" à la main . . . 25
2.6 Illustration du fonctionnement du code pour calculer la distance de Levenshtein . 26
2.7 Résultat renvoyé par le code de la distance de Levenshtein . . . 26
2.8 Extrait du fichier permettant de déterminer les patrons morphosyntaxiques en français . . . 26
2.9 Structure morphosyntaxique relevée pour les termes monolexicaux . . . 27
2.10 Schéma de construction de règle morphosyntaxique . . . 27
2.11 Autre schéma de construction de règle morphosyntaxique . . . 27
2.12 Extrait du troisième script (la clef se situe à la première ligne de chaque if) . . . 27
2.13 Erreurs d'étiquetage produites . . . 28
2.14 Autres erreurs d'étiquetage produites . . . 28
4.1 Calcul des distances de Levenshtein (mots français à gauche, mots anglais à droite) 33
4.2 Schémas des patrons morphosyntaxiques en français (monolexicaux, polylexicaux de taille 2, 3, 4) . . . 36
4.3 Schémas des patrons morphosyntaxiques en français (polylexicaux de taille 5 et 6) 36
4.4 Schémas des patrons morphosyntaxiques en anglais (monolexicaux, polylexicaux de taille 2 et 3) . . . 36
A.1 Annexe : Premier script . . . 47
A.2 Annexe : Deuxième script (partie 1/3) . . . 48
A.3 Annexe : Deuxième script (partie 2/3) . . . 48
A.4 Annexe : Deuxième script (partie 3/3) . . . 48
A.5 Annexe : Troisième script (partie 1/5) . . . 49
A.6 Annexe : Troisième script (partie 2/5) . . . 49
A.7 Annexe : Troisième script (partie 3/5) . . . 49
A.8 Annexe : Troisième script (partie 4/5) . . . 50
A.9 Annexe : Troisième script (partie 5/5) . . . 50
A.10 Annexe : Quatrième script (partie 1/3) . . . 50
A.11 Annexe : Quatrième script (partie 2/3) . . . 51

A.12 Annexe : Quatrième script (partie 3/3) 51

Liste des tableaux

4.1 Variantes graphiques et orthographiques en français et anglais 41
4.2 Suite des variantes graphiques et orthographiques en français et anglais 41
4.3 Variantes morphosyntaxiques en français et anglais 41
4.4 Variantes sémantiques en français et anglais 42

RÉSUMÉ

L'objectif de ce mémoire est de mettre en place une méthode qui permette d'extraire les différents nommages d'ingrédients et actifs cosmétiques à partir de commentaires postés sur le Web en français et en anglais. Les commentaires n'étant pas normalisés, les diverses expressions qui s'y trouvent, amènent à s'intéresser à leurs variations graphiques et morphosyntaxiques. Cette méthode s'inscrit dans un contexte d'acquisition de connaissances et les candidats termes validés pourront être intégrés aux ressources linguistiques et réutilisés dans des analyses ultérieures.

Remerciements

Je remercie toutes les personnes qui m'ont portée lors de l'écriture de ce mémoire. Marguerite Leenhardt pour son aide dans la recherche du sujet et ses corrections apportées, Mathieu Valette pour ses bons conseils pour l'état de l'art et ses corrections apportées également ainsi que mes proches pour leur soutien moteur.

Je tiens aussi à remercier mes collègues de travail qui ont permis à mon stage de bien se dérouler et un grand merci à toute l'équipe ER-TIM pour ces deux années d'enseignement en Master TAL appréciées et au cours desquelles j'ai pu acquérir des compétences en informatique qui guideront mon avenir professionnel.

INTRODUCTION

Ce mémoire a été rédigé dans le contexte du stage de fin d'études ayant eu lieu chez XiKO, une startup spécialisée dans le traitement de commentaires issus du Web (forums, blogs, réseaux sociaux) sur des produits cosmétiques et lifestyle dans le but de faire des analyses et permettre aux clients de mieux comprendre les attentes des consommateurs afin de prendre des décisions marketing par la suite.

Un des projets sur lesquels nous avons travaillés a consisté à déterminer quels ingrédients, parmi une liste prédéfinie, sont les plus utilisés pour des problèmes de peau donnés. L'une des principales tâches à accomplir est donc de capter les noms d'ingrédients dans les commentaires pour pouvoir évaluer ensuite la quantité de messages associant l'un des ingrédients à l'un des problèmes de peau.

Cependant, il est primordial de noter qu'un commentaire n'est pas une entité textuelle normalisée. Les internautes sont libres de citer les ingrédients comme bon leur semble, par des expressions différentes, des raccourcis ou encore avec des fautes d'orthographe ou tout simplement de frappe. Voici quelques exemples pour illustrer ce fait.

Sur la capture d'écran ci-dessous nous pouvons constater qu'en recherchant le motif "rose" dans le corpus de commentaires (via une requête SQL [1]), diverses expressions ont été utilisées par les internautes pour désigner le même ingrédient à savoir l'eau florale de rose comme "hydrolat de rose" ou bien "eau de rose" sans l'adjectif "florale".

Row	left_context	match	
1	bah j'étais déjà fan de ton nettoyant à base de rhassoul et d hydrolat de	rose	, je vais tester celui-ci maintenant merci réo
2	donc ca serait sympas de me conseiller une creme pour tout les jours et jai envi de tester l'eau de	rose	si vous avez une marque potable aussi a me consei
3	s je ne risque rien ? car je me disai quand bronzant un peu on verrai mois ces cicatrices marron et	rose	.
4	j'ai acheté il y'a une semaine l'huile harmonisante à la	rose	musquée de weleda et j'aimerai savoir si ce produit
5	difference. concrètement ça donne : demaquillage a l'eau, nettoyage savon doux, un pschitt d eau de	rose	et si ça tiraille 1 goutte d'huile de jojoba étalée sur l
6	et les gros	rose	quand je touche , je sens un kyste enfin une boule
7	sont sensibles au rancissement (comme la	rose	ou le germe de blé), mais là pas besoin !

Table JSON First < Prev Rows 1 - 7 of 168 Next > Last

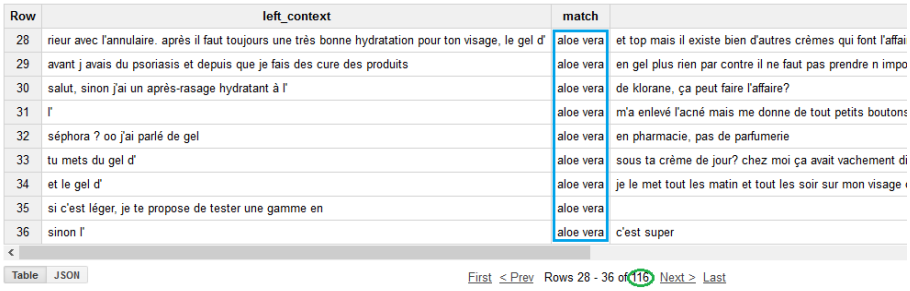
FIGURE 0.1 – Résultat d'une requête SQL demandant d'afficher les contextes gauche et droit du motif "rose"

L'exemple suivant montre la possibilité de trouver un nom d'ingrédient employé de manière raccourcie comme "argan" pour l'huile d'argan.

FIGURE 0.2 – Exemple d'expression raccourcie d'un ingrédient

1. SQL, Structured Query Language, est un langage de programmation conçu pour interroger une base de données au travers de "requêtes".

L'exemple de l'aloe vera, quant à lui, démontre l'importance de la prise en compte des variations graphiques d'une unité lexicale. Il n'est pas rare de voir des fautes de frappe ou d'orthographe dans les messages postés sur Internet. Si on ne les inclut pas dans le système de détection, beaucoup d'occurrences du motif recherché ne seront pas comptabilisées. Ici, la requête porte sur "aloe vera", correctement écrit. Celle-ci donne 116 résultats (cf. information entourée en vert ci-dessous).

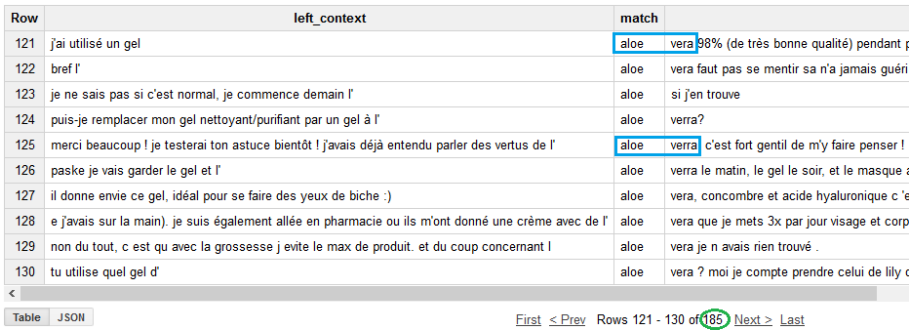


Row	left_context	match	
28	rieur avec l'annulaire. après il faut toujours une très bonne hydratation pour ton visage, le gel d'	aloe vera	et top mais il existe bien d'autres crèmes qui font l'affai
29	avant j avais du psoriasis et depuis que je fais des cure des produits	aloe vera	en gel plus rien par contre il ne faut pas prendre n impo
30	salut, sinon j'ai un après-rasage hydratant à l'	aloe vera	de klorane, ça peut faire l'affaire?
31	l'	aloe vera	m'a enlevé l'acné mais me donne de tout petits boutons
32	séphora ? oo j'ai parlé de gel	aloe vera	en pharmacie, pas de parfumerie
33	tu mets du gel d'	aloe vera	sous ta crème de jour? chez moi ça avait vachement di
34	et le gel d'	aloe vera	je le met tout les matin et tout les soir sur mon visage
35	si c'est léger, je te propose de tester une gamme en	aloe vera	
36	sinon l'	aloe vera	c'est super

Table JSON First < Prev Rows 28 - 36 of 116 Next > Last

FIGURE 0.3 – Résultat d'une requête SQL pour afficher les messages comportant le motif "aloe vera" correctement orthographié

Alors que si nous entrons le motif "aloe" tout court, nous pouvons constater que le nombre de résultats renvoyés augmente jusqu'à 185 et que dans cet exemple précis, il arrive que le nom de cet ingrédient soit écrit avec deux "r" donc de manière erronée mais à prendre en considération.



Row	left_context	match	
121	j'ai utilisé un gel	aloe	vera 98% (de très bonne qualité) pendant
122	bref l'	aloe	vera faut pas se mentir sa n'a jamais guéri
123	je ne sais pas si c'est normal, je commence demain l'	aloe	si j'en trouve
124	puis-je remplacer mon gel nettoyant/purifiant par un gel à l'	aloe	verra?
125	merci beaucoup ! je testerai ton astuce bientôt ! j'avais déjà entendu parler des vertus de l'	aloe	verra c'est fort gentil de m'y faire penser !
126	paske je vais garder le gel et l'	aloe	verra le matin, le gel le soir, et le masque
127	il donne envie ce gel, idéal pour se faire des yeux de biche :)	aloe	vera, concombre et acide hyaluronique c 'e
128	e j'avais sur la main). je suis également allée en pharmacie ou ils m'ont donné une crème avec de l'	aloe	vera que je mets 3x par jour visage et corp
129	non du tout, c est qu avec la grossesse j evite le max de produit. et du coup concernant l	aloe	vera je n avais rien trouvé .
130	tu utilise quel gel d'	aloe	vera ? moi je compte prendre celui de lily

Table JSON First < Prev Rows 121 - 130 of 185 Next > Last

FIGURE 0.4 – Résultat d'une requête SQL pour afficher les messages comportant le motif "aloe"

Avec l'outil implémenté par XiKO pour détecter les motifs recherchés, il est possible d'utiliser les expressions régulières pour capter les éventuelles variantes, ce qui constitue un véritable avantage pour couvrir au mieux toutes les occurrences des ingrédients à condition de parvenir à les identifier en amont. Voici des extraits des ressources linguistiques répertoriées pour le cas précis des ingrédients. Sur la capture d'écran qui suit, nous pouvons voir les variations orthographiques prises en compte pour matcher l'ingrédient "amande douce". L'adjectif "douce" peut être présent ou non et dans certains cas il y a une lettre manquante ou supplémentaire.

```
<group id="almond">
    <form>amande (dd?o?u?ce)?</form>
</group>
```

FIGURE 0.5 – Extrait de ressources linguistiques - amande douce

Le nombre est aussi une caractéristique à ne pas omettre car le motif peut être mis au singulier ou au pluriel par l'internaute et ne pas être détecté par le système si cette éventualité n'est pas spécifiée.

```
<group id="butter">
    <form>beurr?es?</form>
</group>
```

FIGURE 0.6 – Extrait de ressources linguistiques - beurre

Les synonymes sont courants donc il est important de les repérer. Sur la capture d'écran ci-dessous, nous pouvons voir les différentes dénominations relevées pour l'ingrédient "AHA" (Alpha Hydroxy Acid en anglais) qui n'est autre que l'acide alpha hydroxylé ou encore l'acide glycolique.

```
<group id="AHA">
    <form>acides? (α|alpha)[- ]hydroxylés?</form>
    <form>A-?HA</form>
    <form>acides? glycoliques?</form>
    <form>acides? hydroxyacétiques?</form>
    <form>peeling chimique</form>
</group>
```

FIGURE 0.7 – Extrait de ressources linguistiques - AHA

Il est également possible de rencontrer des abréviations, qui sont à considérer comme des synonymes. Ici, nous avons "hv" pour indiquer "huile végétale".

```
<group id="oil">
    <form>huiles?</form>
    <form>hv</form>
</group>
```

FIGURE 0.8 – Extrait de ressources linguistiques - huile

Ces observations nous amènent à la problématique de ce mémoire formulée ainsi : "**Acquisition de connaissances à des fins d'analyse automatique : Extraction des différentes façons de nommer les ingrédients et actifs cosmétiques dans les conversations spontanées des internautes en français et anglais et comparaison multilingue.**"

Il s'agit de tenter de trouver une méthode automatique qui puisse identifier toutes les variantes existantes de nommage des ingrédients ciblés dans un ensemble de commentaires afin d'alimenter la base de connaissances de l'entreprise.

La première partie se compose d'abord d'un état de l'art axé sur l'extraction terminologique, qui donne une mise en contexte et présente les méthodologies de quelques travaux antérieurs menés à cette fin nous permettant d'avoir des pistes de réflexion pour entamer notre extraction. S'ensuit la présentation des méthodes expérimentées pour parvenir à répondre à notre problématique. La deuxième partie est constituée d'une description des corpus utilisés ainsi que des résultats obtenus pour chacune des deux langues et leur comparaison avec une conclusion générale pour donner le mot de la fin.

Première partie

Contexte général

CHAPITRE 1

ÉTAT DE L'ART

1.1 Introduction

Le sujet de ce mémoire se rapproche beaucoup de l'extraction terminologique. Cette application du TAL consiste à extraire les termes potentiels d'un corpus spécialisé de façon automatique grâce à un "extracteur de termes". Dans notre cas, les ingrédients et actifs cosmétiques peuvent être considérés comme des concepts et leurs différents nommages comme les termes à extraire.

Comme l'explique [L'Homme, 2004], l'extraction terminologique a fait partie du processus d'intégration des outils informatiques à la terminographie [1] au cours des années 80 donnant lieu à la terminotique. Les terminographes ont ainsi peu à peu commencé à entreprendre différemment la création de ressources terminologiques telles que les dictionnaires spécialisés, banques terminologiques et listes de termes destinées aux rédacteurs ou traducteurs. L'extraction terminologique a fait l'objet de nombreuses recherches tout au long des années 90. TERMINO ouvre la marche des extracteurs de termes en français. Il a été suivi par de nombreux autres dont les plus connus sont ACABIT [Daille, 1994a, Daille, 1994b], LEXTER[Bourigault, 1994, Bourigault et al., 1996], FASTR [Jacquemin, 1997] que [Cabré et al., 2001] présente brièvement, ou encore YaTeA [Aubin and Hamon, 2006] et TTC TermSuite [Morin and Daille, 2012] plus récents. Désormais, l'extraction terminologique sert pour toutes sortes d'activités comme la construction de glossaires, ontologies, index et trouve sa place dans le monde de l'entreprise à travers le besoin d'analyse abondant des données textuelles.

1.2 Définition et principe

[L'Homme, 2004] explique que ce qui différencie un terme des autres unités lexicales est son sens spécialisé par rapport à un domaine de spécialité et précise qu'il existe deux types de termes : les termes simples (une seule entité graphique) et les termes complexes (plusieurs entités graphiques). En plus des critères morphologiques et syntaxiques dans leur définition respective du terme, [Daille, 1994b] indique que son critère sémantique implique une référence à un concept unique et [Ibekwe-SanJuan, 2001] ajoute qu'un terme doit être interprétable hors contexte. Il en ressort qu'un terme est un mot ou une séquence de mots qui désigne les concepts d'un domaine spécialisé. De nombreux extracteurs de termes sont conçus

1. [Rey, 1992] désigne par "terminographie" la partie pratique de la terminologie.

pour détecter les syntagmes nominaux car il s'agit de la forme des termes la plus répandue.

Pour [Korenchuk, 2014] l'objectif des systèmes d'extraction terminologique consiste à extraire les unités lexicales susceptibles d'être des termes. Cela nous permet d'introduire la notion fondamentale de "candidat terme" car à l'issue de leurs analyses du corpus, suivant les règles qu'on leur a définies, les extracteurs de termes proposent un ensemble de mots ou suites de mots pouvant être des termes. Ces candidats termes devront ensuite être validés par un expert du domaine en question afin d'alimenter les diverses ressources terminologiques.

La production d'un extracteur de termes ne peut donc être entièrement exacte. Après la vérification nécessaire, les séquences non admises comme termes seront considérées comme du "bruit" et à l'inverse, les véritables termes qui n'auront pas été relevés généreront du "silence". Le bruit et le silence diminuent respectivement le taux de précision et de rappel sur lesquels repose l'évaluation des extracteurs de termes.

1.3 Différentes méthodes

Il existe trois approches pour l'extraction terminologique. La première est linguistique, elle consiste à utiliser de différentes façons les informations grammaticales des mots comme les parties du discours (Part Of Speech) pour détecter les candidats termes. La deuxième est statistique car elle s'appuie sur divers calculs basés sur la fréquence, la répartition, l'association ou encore la distance entre les unités lexicales pour filtrer les candidats termes. [Drouin and Langlais, 2006] présente et teste quelques unes de ces mesures pour évaluer le potentiel terminologique de candidats termes. La troisième est appelée "hybride" ou "mixte" car elle allie des éléments de l'approche linguistique et des éléments de l'approche statistique.

L'étiquetage du corpus en parties du discours est souvent la première étape pour l'analyse linguistique afin d'exploiter les catégories grammaticales des mots comme repères. [Daille, 1994b] les utilise pour construire des patrons morphosyntaxiques correspondants aux syntagmes nominaux relevés manuellement.

Elle constate que les termes sont majoritairement constitués de deux éléments principaux [2] et ceux qui en comportent plus sont le résultat d'une surcomposition, modification ou coordination des séquences plus courtes. L'extracteur de termes ACABIT que [Daille, 1994b] développe se focalise donc sur la détection des termes binaires selon les patrons morphosyntaxiques identifiés précédemment. L'unique séquence de plus trois éléments principaux prise en compte est celle qui contient une insertion adjectivale.

Pour démarrer l'extraction des noms des ingrédients et actifs cosmétiques, l'utilisation des patrons morphosyntaxiques est une piste intéressante. Elle peut servir à relever les structures des syntagmes nominaux de manière contrôlée.

[Bourigault et al., 1996] exploite les parties du discours de manière inversée : le module de découpage de l'extracteur de termes LEXTER identifie d'abord les mots

2. [Daille, 1994b] ne considère pas les prépositions et les déterminants comme éléments principaux.

dont les étiquettes grammaticales ont été fixées comme n'appartenant pas à un terme (verbe, pronom, préposition + article possessif, etc) afin de définir les frontières des termes potentiels. En cas d'ambiguïté, le système d'apprentissage endogène basé sur le corpus va trancher en fonction des cas étudiés pendant la première analyse. A la fin de cette étape, des "syntagmes nominaux de longueur maximale" sont délimités à partir desquels un autre module extrait les candidats termes en respectant le principe de dérivation syntagmatique qui établit qu'un terme complexe est formé d'une séquence principale (appelée "tête") et d'une séquence attribut (appelée "expansion"). Si la tête ou l'expansion est constituée de plusieurs mots, ceux-ci doivent suivre le même schéma entre eux. Le système d'apprentissage entre à nouveau en jeu si deux possibilités d'extraction se concurrencent.

Cette méthode présente une autre façon d'utiliser les catégories grammaticales. Déterminer les frontières des candidats termes semble être moins pertinent pour le sujet de ce mémoire car nous disposons d'éléments de base au préalable pour orienter la détection des expressions des ingrédients et actifs cosmétiques dans les commentaires. Cependant, la mise en place de règles pour aiguiller l'extraction des termes est à prendre en compte.

[Ibekwe-SanJuan, 2001] s'appuie sur les étiquettes grammaticales pour construire des automates sur le logiciel INTEX. Ces derniers extraient successivement des syntagmes nominaux maximaux (SN-max), moyens (SN-moy) et minimaux (SN-min). Les syntagmes nominaux minimaux sont des candidats termes. Selon la logique adoptée, un SN-max comprend des SN-min reliés par des prépositions et/ou des marqueurs de coordination. Si un SN-max présente une coordination équilibrée, ce qui veut dire que les unités qui entourent les éléments de la coordination sont indépendants comme dans une énumération par exemple, l'extraction de SN-min ne pose pas de problème. Si la coordination est non équilibrée, des erreurs d'extraction sont générées car cela demande une analyse complexe. Un syntagme nominal maximal sans coordination devient un syntagme nominal moyen, analysé en prenant les prépositions comme repères pour extraire les SN-min restants.

Dans cette méthodologie, l'intérêt porté à la coordination est pertinent car ce phénomène est régulièrement rencontré et nécessite un traitement particulier pour en extraire correctement les termes. La distinction faite entre coordination équilibrée et non équilibrée permet de mieux comprendre l'enjeu présenté par cette structure linguistique dans le cadre de l'extraction terminologique.

[Korenchuk, 2014] entreprend une toute autre démarche en mettant à profit au maximum le corpus afin de générer des ressources endogènes et avoir moins recours aux ressources externes. L'algorithme de [Vergne, 2003, Vergne, 2004] est appliqué pour discerner les mots informatifs (mots longs peu fréquents entourés de mots courts plus fréquents) de ceux qui ne le sont pas (mots vides). Les mots annotés comme informatifs dans 90% des cas sont ensuite filtrés par le calcul de la fréquence absolue d'une part et du TF IDF (version adaptée au corpus) d'autre part. Les unités lexicales arrivées en tête de listes sont alors retenues comme candidats termes monolexicaux et vont servir à la création de ressources morphologiques afin d'identifier d'autres candidats termes monolexicaux moins fréquents. Le procédé mis en oeuvre consiste à faire ressortir leurs morphèmes caractéristiques (tri-grammes de début et milieu de mot) pour constituer une liste de formants. Celle-ci sert ensuite à calculer un score de confiance pour chaque mot analysé avec un point attribué pour

chaque formant contenu. Les mots ayant accumulé trois points ou plus sont ajoutés aux candidats termes monolexicaux relevés. Cette étape débouche sur l'extraction des candidats termes polylexicaux suivant des patrons morphosyntaxiques endogènes en respectant une série de règles de structuration précises. Les tri-grammes en fin de mots informatifs sont utilisés comme des étiquettes morphosyntaxiques. Les candidats termes polylexicaux extraits ne sont pas limités aux syntagmes nominaux. Ceux qui apparaissent dans le premier tiers sont relevés et pour améliorer les résultats, il est établit qu'au moins l'un des composants doit être un candidat terme monolexical.

Cette méthode montre une manière différente d'aborder l'extraction terminologique. La construction de ressources endogènes amène à examiner les mots et leurs formes en profondeur, ce qui permet d'envisager de nouvelles pistes. La dernière condition formulée peut être appropriée pour l'extraction des ingrédients et actifs cosmétiques.

La méthode proposée par [Korenchuk, 2014] suit l'approche hybride car elle associe des analyses linguistiques et des mesures statistiques. C'est également le cas de la méthode appliquée par [Daille, 1994b], après avoir relevé les candidats termes selon les patrons morphosyntaxiques, ACABIT fait des calculs statistiques pour les filtrer avec la mesure du Loglike [Dunning, 1993], qui donne une meilleure sélection de candidats termes que la fréquence seule [3]. La méthode de [Bourigault et al., 1996] ainsi que celle de [Ibekwe-SanJuan, 2001] gardent une approche linguistique. Pour cette dernière, le filtrage des candidats termes par fréquence éliminerait beaucoup de termes potentiels moins fréquents dans le corpus dont certains sont des hapax.

Pour l'extraction des différentes façons de nommer des ingrédients et actifs cosmétiques, cette constatation est pertinente car la fréquence ne constitue pas un indice significatif. Pour autant, l'approche hybride n'est pas écartée de notre travail. D'autres mesures sont à exploiter comme celles de la distance jugées intéressantes par [Daille, 1994b] en ce qui concerne la variation des termes.

LEXTER [Bourigault et al., 1996] présente les candidats termes extraits sous la forme d'un réseau terminologique, ce qui est propice à la visualisation des variantes terminologiques. Construit par le module de structuration, il regroupe les candidats termes par leurs éléments communs (tête ou expansion). Cette organisation est effectuée sur la base d'un coefficient qui permet de voir combien un terme est productif. Plus il fait partie d'un grand nombre de candidats termes, plus il est probable qu'il soit un terme validé et que ceux qui sont formés avec le soient aussi.

Dans leur article respectif, [Daille, 1994b] et [Tartier, 2006] présentent une typologie de la variation terminologique. Ces indications permettent d'avoir une vue d'ensemble sur les éléments de variation de termes possibles.

[Daille, 1994b] distingue trois grands types de variantes avec les variantes graphiques et orthographiques (présence ou non d'une majuscule, trait d'union ou marque du pluriel), les variantes morphosyntaxiques (modification de la structure du terme par la présence ou non d'un article ou d'une préposition, synonymie de deux termes de structures différentes) et les variantes elliptiques (manifestation d'un terme de deux éléments principaux par un seul des deux).

[Tartier, 2006] différencie la variation morphologique (dérivation, composition), et syntaxique (changements de structure minimes, présence ou ellipse d'un nom "support", insertion, expansion ou substitution d'un élément du terme) en plus de la varia-

3. Par rapport à d'autres mesures que [Daille, 1994b] teste.

tion morphosyntaxique. L'auteur mentionne également la variation orthographique, sémantique (synonymes, hyperonymes) ainsi que les sigles et abréviations comme autres formes de variation.

Cette organisation et classifications sont utiles pour notre travail d'extraction car les variantes terminologiques en sont le point central.

CHAPITRE 2

MÉTHODES

2.1 Introduction

Notre objectif consiste à extraire les différentes expressions présentes dans les commentaires pour désigner les ingrédients et actifs cosmétiques à partir d'une liste d'éléments de base [1].

Ces éléments de base sont des noms d'ingrédients sous forme minimale qui vont servir d'ancre dans le processus de détection et d'extraction. Il s'agit uniquement d'unités monolexicales pour simplifier les traitements automatiques qui suivent et par la même occasion ne pas influer sur la structure des candidats termes extraits. Dans le cas d'un ingrédient ou d'un actif dont le nom est composé de plusieurs mots comme "aloe vera", "argile verte" ou encore "fleur d'oranger", l'unité de sens retenue est celle qui en est la plus représentative en contexte. Donc, "aloe", "argile" et "oranger" sont considérés comme les éléments de base pour ces cas précis.

Les listes des éléments sélectionnés pour le français et l'anglais sont les suivantes :

```
["chanvre", "cacao", "jojoba", "amande", "argan", "karité", "sésame", "noisette",
 "hamamélis", "bleuet", "oranger", "lavande", "menthe", "rose",
 "hyaluronique", "propolis", "aloe", "argile", "aha", "acide"]
```

FIGURE 2.1 – Liste des éléments de base en français

```
["hemp", "cacao", "jojoba", "almond", "argan", "shea", "sesame", "hazelnut",
 "hamamelis", "hazel", "cornflower", "orange", "lavender", "mint", "rose",
 "hyaluronic", "propolis", "aloe", "clay", "aha", "acid"]
```

FIGURE 2.2 – Liste des éléments de base en anglais

La seconde liste compte un élément de plus que la précédente car elle contient un autre nom courant d'hamamélis en anglais, "witch hazel", entré "hazel". Dans les deux listes, le dernier élément inclus est "acid(e)" pour pouvoir relever les éventuels synonymes de "AHA" (cf. introduction générale), ce qui n'est pas possible avec seulement cet acronyme dans les listes.

Notre méthodologie met en oeuvre une approche linguistique pour détecter les variations morphosyntaxiques, associée à un calcul statistique pour prendre en compte les variantes graphiques et orthographiques des éléments de base.

1. Cette logique rejoint celle adoptée pour l'extracteur de termes FASTR [Jacquemin, 1997] qui part aussi d'éléments connus à savoir d'autres termes pour extraire leurs variantes.

Toutes les étapes sont appliquées deux fois. D'abord sur le corpus en français puis sur le corpus en anglais. Elles reposent sur des scripts codés en Python 3 et des commandes Bash pour trier les résultats obtenus et éliminer leurs doublons à chaque avancée.

2.2 Etape 1 : Etiquetage morphosyntaxique

Tout d'abord, nous procédons à l'étiquetage des corpus en catégories grammaticales afin de constituer une base de travail exploitable. Pour cela nous lançons Treetagger en ligne de commande avec les paramètres adéquats pour chacune des langues. Les fichiers en sortie sont des .txt qui contiennent pour chaque token du corpus analysé, sa forme, sa partie du discours (POS) et son lemme. Les POS vont nous permettre de déterminer les patrons morphosyntaxiques des diverses expressions des ingrédients et actifs cosmétiques dans les commentaires pour pouvoir ensuite détecter ces derniers automatiquement.

Après vérification des étiquetages effectués, nous constatons qu'en français, certains des éléments de base ne sont pas reconnus par Treetagger, ce qui entraîne une annotation erronée : "jojoba", "oranger" et "vera" de "aloe vera" sont étiquetés en tant que verbes. Si nous ignorons ces erreurs, les patrons morphosyntaxiques seront faussés et généreront du bruit. Il faut donc changer [2] ces étiquettes par celles qui conviennent mieux : NOM pour "jojoba" et "oranger" et "ADJ" (adjectif) pour "vera". Certains d'entres eux, ont également leur lemme marqué "<unknown>", ce qui rend cette information inexploitable.

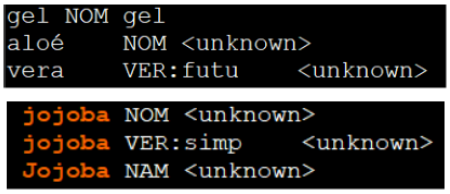


FIGURE 2.3 – Étiquetage d'éléments de base en français

En anglais, ces mêmes éléments sont reconnus et sont bien étiquetés sauf "rose" annoté aussi comme un verbe à cause de son ambiguïté sémantique (prétérit de "rise", verbe irrégulier). Nous changeons donc son POS par "NN" (nom commun). Son lemme n'est donc pas correct non plus. Pour les autres éléments, il n'y a pas d'erreur constatée à ce niveau.

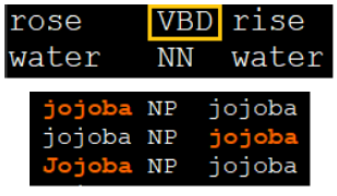


FIGURE 2.4 – Étiquetage d'éléments de base en anglais

L'étiquetage en catégories grammaticales rétablie dans les deux langues, nous pouvons procéder à l'implémentation du système de détection et d'extraction fondé

2. Via la fonction "remplacer" de l'éditeur de texte Notepad++

sur la correspondance des formes des tokens avec les éléments de base.

2.3 Etape 2 : Calcul des distances entre les mots

Nous devons ensuite procéder à l'élargissement des listes des éléments de base à leurs variantes graphiques existantes dans les corpus pour pouvoir prendre en compte toutes les formes.

La présente étape nous permet d'y parvenir grâce à la mesure de la distance de Levenshtein [Levenshtein, 1966], aussi appelée distance d'édition. Son algorithme renvoie un chiffre correspondant à la distance entre deux chaînes de caractères. Plus la distance est grande, plus les mots comparés sont graphiquement éloignés. Ce chiffre est calculé en fonction de trois paramètres : la suppression, l'insertion et la substitution d'un caractère avec un point attribué pour chacun de ces cas.

Il est possible de faire ce calcul à la main, en voici un exemple : la distance entre les deux chaînes de caractères "argan" (élément de base) et "argane" (forme existante dans le corpus français) est 1. Une modification est visible, l'insertion du caractère "e" à la fin de la forme provenant d'un commentaire. Ce résultat permet de voir que cette unité lexicale est proche de l'élément de base comparé et est donc susceptible d'en être une variante.

			a	r	g	a	n
		i=0	i=1	i=2	i=3	i=4	i=5
	j=0	0	1	2	3	4	5
a	j=1	1	0	1	2	3	4
r	j=2	2	1	0	1	2	3
g	j=3	3	2	1	0	1	2
a	j=4	4	3	2	1	0	1
n	j=5	5	4	3	2	1	0
e	j=6	6	5	4	3	2	1

FIGURE 2.5 – Calcul de la distance de Levenshtein entre "argan" et "argane" à la main

Cette mesure est donc un bon moyen pour identifier les unités lexicales qui sont graphiquement proches des éléments de base dans les corpus. Nous intégrons à notre premier script une fonction calculant cette distance (donnée comme exemple sur un site Internet [3]). Ainsi, tous les tokens ayant obtenus une distance inférieure ou égale à 2 sont relevés par notre programme (cf. annexe). Au delà de ce seuil, cela ramène trop de bruit.

3. Lien du site : https ://www.programcreek.com/python/example/94974/Levenshtein.distance

Une fois les variantes triées, nous les ajoutons aux listes des éléments de base respectives pour le français et l'anglais.

```
def levenshtein_distance(a, b):
    """Return the Levenshtein edit distance between two strings *a* and *b*."""
    if a == b:
        return 0
    if len(a) < len(b):
        a, b = b, a
    if not a:
        return len(b)
    previous_row = range(len(b) + 1)
    for i, column1 in enumerate(a):
        current_row = [i + 1]
        for j, column2 in enumerate(b):
            insertions = previous_row[j + 1] + 1
            deletions = current_row[j] + 1
            substitutions = previous_row[j] + (column1 != column2)
            current_row.append(min(insertions, deletions, substitutions))
        previous_row = current_row
    return previous_row[-1]


###
mot1 = "argan"
mot2 = "argane"

dist = levenshtein_distance(mot1, mot2)

print ("Distance de Levenshtein entre " + mot1 + " et " + mot2 + " : " + str(dist))
```

FIGURE 2.6 – Illustration du fonctionnement du code pour calculer la distance de Levenshtein

```
Distance de Levenshtein entre argan et argane : 1
```

FIGURE 2.7 – Résultat renvoyé par le code de la distance de Levenshtein

2.4 Etape 3 : Recherche des patrons morphosyntaxiques

Dans cette étape, un deuxième script est conçu pour créer un fichier .csv permettant de répertorier manuellement les patrons morphosyntaxiques des différents nommages des ingrédients et actifs cosmétiques dans les commentaires.

Pour se faire, à chaque élément de base identifié, notre programme le relève avec son contexte gauche et droit (fenêtre de 6 tokens de part et d'autre) ainsi que leurs parties du discours associées.

Puis nous ouvrons ce nouveau fichier dans un tableur afin de procéder au prélèvement méthodique des patrons morphosyntaxiques recherchés dans le premier tiers du document.

Voici une capture d'écran qui donne un aperçu du déroulement de cette opération :

phytothérapie	NOM	L'	DET:ART	Aloe	NAM	Vera	NAM
huile	NOM	de	PRP	rose	NOM	musquée	ADJ
huile	NOM	d'	PRP	argon	NOM	,	PUN
de	PRP	l'	NOM	acide	ADJ	hyaluronique	ADJ
l'	NOM	acide	ADJ	hyaluronique	ADJ	.	SENT
huile	NOM	d'	PRP	argane	NOM	pure	ADJ

FIGURE 2.8 – Extrait du fichier permettant de déterminer les patrons morphosyntaxiques en français

Les termes monolexicaux constituent un cas à part étant donné qu'il s'agit que d'une seule unité lexicale. Pour rendre leur extraction possible, nous avons décidé de les relever s'ils sont précédés d'un déterminant. Cela permet de les analyser comme des termes polylexicaux de deux éléments avec une structure morphosyntaxique adaptée.

l'	DET:ART	aloès	NOM	avec	PRP
non	ADV	acide	ADJ	.	SENT
l'	DET:ART	aloe	NOM	.	SENT

FIGURE 2.9 – Structure morphosyntaxique relevée pour les termes monolexicaux

2.5 Etape 4 : Implémentation des règles des patrons morphosyntaxiques

Il faut maintenant implémenter les patrons morphosyntaxiques identifiés sous forme de règles dans le troisième script pour que celui-ci extrait toutes les suites de mots leurs correspondants dans le corpus de la langue en cours de traitement, toujours avec un élément de base comme repère.

Pour réaliser cela, deux tâches sont nécessaires à accomplir en amont. La première consiste à ranger les patrons morphosyntaxiques par longueur, c'est-à-dire le nombre d'éléments qui les compose (monolexicaux et polylexicaux de taille variée). La seconde consiste à marquer la position de l'élément de base dans chacune des structures morphosyntaxiques afin de la considérer comme clef et construire les règles d'extraction autour d'elle. Voici deux exemples ainsi qu'une capture d'écran du script pour illustrer cela :

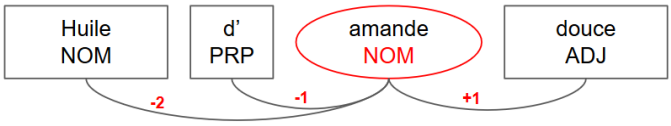


FIGURE 2.10 – Schéma de construction de règle morphosyntaxique

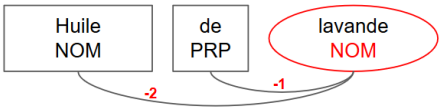


FIGURE 2.11 – Autre schéma de construction de règle morphosyntaxique

```
# Backslash utilisés pour revenir à la ligne pour plus de lisibilité
for p_base in positions_base :

    #monolexicaux
    if (liste2[p_base][1] == "NAM" or liste2[p_base][1] == "NOM") \
       and (liste2[p_base -1][1] == "DET:ART") :
        fichier_pos.write(liste2[p_base -1][0] + " " + liste2[p_base][0] + "\n")

    #polylexicaux : 2
    if (liste2[p_base][1] == "ADJ") \
       and (liste2[p_base +1][1] == "ADJ" or liste2[p_base +1][1] == "NOM" or \
            liste2[p_base +1][1] == "NAM") :
        fichier_pos.write(liste2[p_base][0] + " " + liste2[p_base +1][0] + "\n")
```

FIGURE 2.12 – Extrait du troisième script (la clef se situe à la première ligne de chaque if)

Il est possible de rassembler plusieurs patrons morphosyntaxiques en une règle à condition que la position de l'élément de base (clef) soit la même et qu'un seul POS diffère entre eux.

Pour l'étiquetage en français, Treetagger annote les noms communs "NOM" et lorsque ceux-ci commencent par une majuscule, ils deviennent des noms propres "NAM" (name) alors que ce n'est pas toujours vrai. C'est pourquoi dans les règles instaurées, nous regroupons ces deux parties du discours. Cela signifie que dès qu'il y a un "NOM" dans une règle, un "NAM" peut également y répondre et inversement.

Nous retrouvons un cas similaire dans l'étiquetage en anglais dans la mesure où Treetagger attribue une étiquette grammaticale différente s'il s'agit d'un nom commun au singulier ou au pluriel et la même chose pour les noms propres. C'est pourquoi, dans certaines règles concernées, nous avons inclut la possibilité que le dernier élément de la structure soit un nom commun ou propre au singulier ou au pluriel.

2.6 Etape 5 : Nettoyage des expressions relevées pour livrer les candidats termes

Cette dernière étape est mise en place pour nettoyer les éventuelles petites unités parasites qui se glissent lors de l'extraction des expressions suivant les règles des patrons morphosyntaxiques.

Cela est dû à une mauvaise étiquette grammaticale qui leur est attribuée par Treetagger pour deux raisons. D'une part, il s'agit de corpus de commentaires donc les tokens ne sont pas normalisés. Par exemple des signes de ponctuation peuvent être collés aux mots sans que ce soit la norme ou bien des mots inconnus de l'étiqueteur comme des onomatopés sont susceptibles de l'induire en erreur.

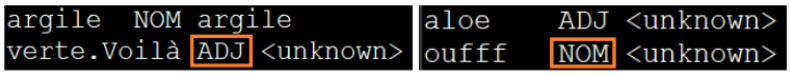


FIGURE 2.13 – Erreurs d'étiquetage produites

D'autre part, Treetagger est parfois lui même la cause de ces erreurs dans le cas où, comme nous l'avons vu dans l'étape précédente, il annote un nom commun commençant par une majuscule comme nom propre ou bien une série de chiffres terminée par une ou plusieurs lettres comme un nom.

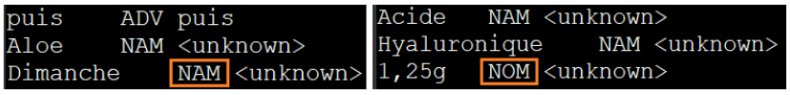


FIGURE 2.14 – Autres erreurs d'étiquetage produites

Pour rendre les candidats termes extraits de meilleure qualité, notre quatrième script retire les signes de ponctuation sauf le trait d'union, la virgule, le point virgule, les deux points et la barre oblique s'ils se trouvent à l'intérieur de l'expression ainsi que les symboles tels que le pourcent, l'astérisque, le symbole euro et le plus sauf si ce dernier est aussi à l'intérieur de l'expression pour ne pas altérer son interprétation. Les chiffres sont également supprimés et les mots "parasites" (cf. résultats des expériences) le sont par le biais d'une "stoplist" constituée à cette effet.

Deuxième partie

Expérimentations

CHAPITRE 3

CORPUS

3.1 Corpus et préparation

Dans une visée multilingue, nous avons pris deux corpus comme objet d'analyse : l'un en français, l'autre en anglais. Tous deux sont de taille similaire (environ 822 000 mots pour le premier et 825 000 mots pour le second). Ils proviennent de corpus de commentaires plus grands, créés pour le besoin des projets menés par XiKO. Les commentaires réunis sont issus de blogs sur le thème de la beauté et d'autres plateformes comme doctissimo.fr, aufeminin.com, magicmaman.com pour le français ou encore amazon.com, walmart.com, asda.com pour l'anglais.

Les métadonnées qui accompagnent les commentaires ne sont pas utiles dans les processus automatiques entrepris dans ce mémoire, c'est pourquoi seuls les contenus des messages sont extraits et mis dans un fichier .txt pour chaque langue.

Afin de faciliter l'étiquetage morphosyntaxique, l'uniformisation de ces fichiers consiste à supprimer les guillemets qui servent à délimiter les messages ainsi que remplacer les petites insertions superflues, entrevues dans le corpus anglais, telles que " ;" (espace) ou "&rsquo ;" (apostrophe) par les caractères qu'elles encodent. Pour finir, il faut s'assurer que l'encodage est en UTF-8.

CHAPITRE 4

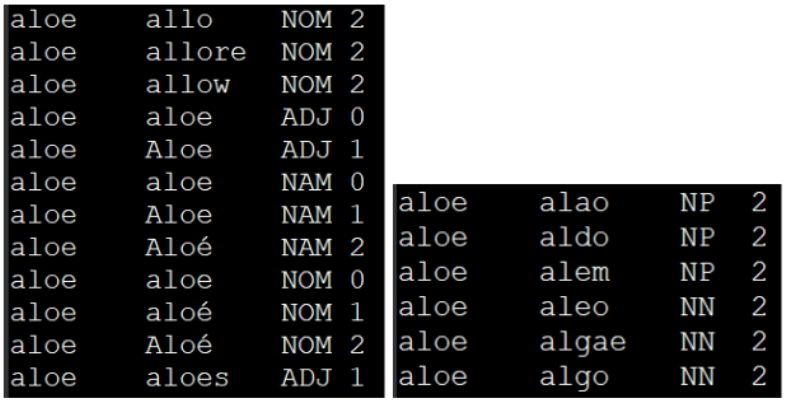
RÉSULTATS

Ce chapitre présente les résultats que nous avons obtenus tout au long du processus d'extraction.

4.1 Variantes des éléments de base

Revenons à l'étape du calcul des distances de Levenshtein pour capter les variantes des éléments de base.

Notre premier script a relevé tous les tokens dont la distance était inférieure ou égale à 2 par rapport aux ingrédients de base : 1409 mots pour le français et 1890 pour l'anglais (totaux sans doublons). Sans surprise, la plupart d'entre eux étaient du bruit comme nous pouvons le voir sur les captures d'écran ci-dessous (lecture des colonnes : élément de base / token / POS du token / distance calculée).



aloe	allo	NOM	2
aloe	allore	NOM	2
aloe	allow	NOM	2
aloe	aloe	ADJ	0
aloe	Aloe	ADJ	1
aloe	aloe	NAM	0
aloe	Aloe	NAM	1
aloe	Aloé	NAM	2
aloe	aloe	NOM	0
aloe	aloé	NOM	1
aloe	Aloé	NOM	2
aloe	aloes	ADJ	1

aloe	alao	NP	2
aloe	aldo	NP	2
aloe	alem	NP	2
aloe	aleo	NN	2
aloe	algae	NN	2
aloe	algo	NN	2

FIGURE 4.1 – Calcul des distances de Levenshtein (mots français à gauche, mots anglais à droite)

Voici les variantes retenues pour chaque langue après leur tri, en prenant soin de vérifier les contextes dans les corpus étiquetés en cas de doute.

Variantes des éléments de base pour le français (total : 60)

acid
Acide
acides
Acides
Aha

aléo
Aloe
aloé
Aloé
aloes
aloès
aloex
Aloex
amandes
ammande
aoloes
Argan
argane
argen
argil
Argile
argiles
argon
argyle
Bleuet
bleut
Cacao
daloe
daloé
damandes
dargan
hamamelis
Hamamelis
hammanelis
hialuronique
hialuroniques
huyaluronique
hyalurinuque
Hyaluronique
hyaluroniques
jojba
Jojba
Jojoba
kariete
karite
Karite
Karité
largil
largile
lavance
Lavande
lavandin
manthe
Menthe

menthol
Noisette
noisettes
Propolis
Rose
roses

Variantes des éléments de base pour l'anglais (total : 17)

Acid
acids
Acids
aleo
Almond
Aloe
Argan
Hazel
jajoba
Jojoba
lavendar
mint-
Mint
Orange
Rose
roses
Shea

Il est important de noter qu'une plus grande diversité des variantes est présente dans les résultats en français qu'en anglais, car le corpus anglais est un échantillon d'un vaste corpus qui n'était pas spécialement dédié à l'analyse des ingrédients et actifs cosmétiques.

Nous pouvons y distinguer les variations graphiques et orthographiques déjà évoquées (majuscules, marque du pluriel, diacritiques, fautes d'orthographes ou de frappe). Comme nous l'avons expliqué plus tôt dans ce mémoire, ces variantes ont été ajoutées aux listes des éléments de base initiaux.

4.2 Patrons morphosyntaxiques

Grâce aux fichiers créés avec le deuxième script, lisibles dans un tableur, nous avons relevé à la main les différents patrons morphosyntaxiques présents parmi 525 lignes pour le français et 230 lignes pour l'anglais (respectivement un tiers des fichiers). Cela nous a permis de constituer ensuite les schémas morphosyntaxiques avec les clefs (marquées en rouge) à suivre pour l'implémentation des règles dans le script suivant. L'anglais en a moins pour les raisons précédemment mentionnées.

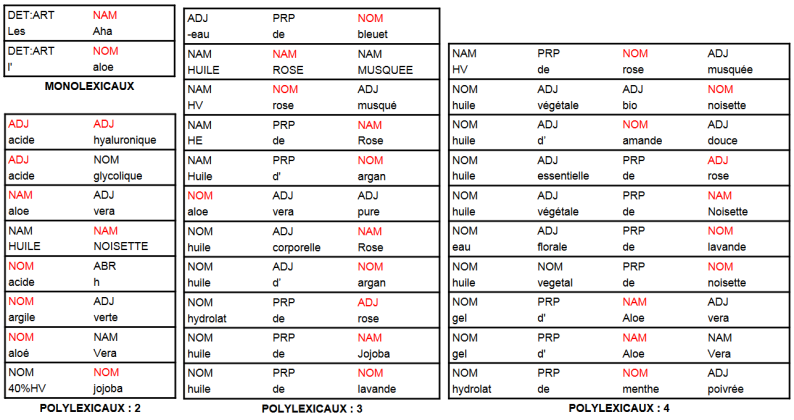


FIGURE 4.2 – Schémas des patrons morphosyntaxiques en français (monolexicaux, polylexicaux de taille 2, 3, 4)

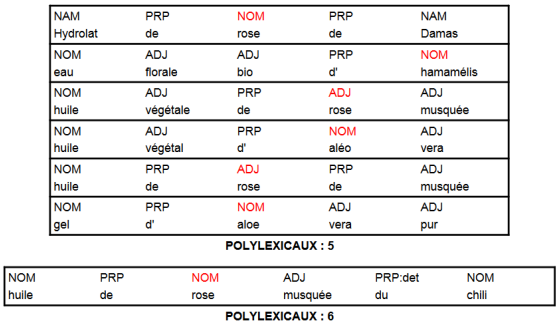


FIGURE 4.3 – Schémas des patrons morphosyntaxiques en français (polylexicaux de taille 5 et 6)

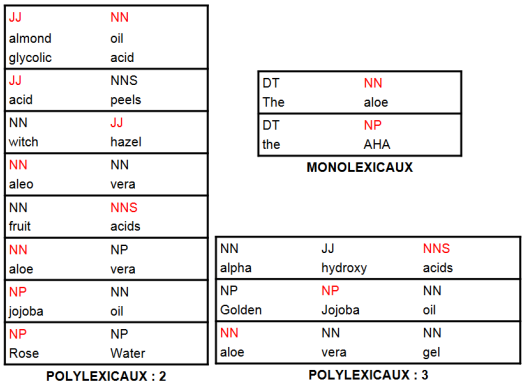


FIGURE 4.4 – Schémas des patrons morphosyntaxiques en anglais (monolexicaux, polylexicaux de taille 2 et 3)

Voici un exemple de la façon dont les règles d'extraction ont été formulées : Prenons le cas des patrons morphosyntaxiques monolexicaux en français, la clef se situe à la même place et seulement un élément diffère entre les deux schémas (la clef en l'occurrence) donc toutes les conditions sont réunies pour rassembler ces deux structures en écrivant une seule règle => Si l'élément clef est un nom propre ou un nom commun, et que l'élément qui le précède est un déterminant, alors extrait cette suite dans l'ordre inverse (déterminant, élément clef).

4.3 Candidats termes

Le troisième script a extrait 678 suites morphosyntaxiques correspondantes aux règles instaurées pour le français et 511 pour l'anglais (totaux sans doublons). En voici quelques unes :

Expressions relevées lors de l'extraction en français :

huile de chanvre J'
huile de chanvre matin
huile de chanvre pure
huile de jojba
huile de jojoba
huile de Jojoba
Huile de Jojoba
huile de jojoba +bois
huile de jojoba personnelement
huile de jojoba Suisse
huile de karité
huile de lavande
Huile de Lavande
huile de noisette
huile de noisette ,je
Huile de noisettes
huile de rose
Huile de Rose
huile de rose de musquée
huile de rose musquée
Huile de Rose Musquée
huile de rose musquée bio
huile de rose musquée du chili
huile de rose naturelle
huile essenssielle de lavande
huile essentiel de lavande

Expressions relevées lors de l'extraction en anglais :

rose water
Rose Water
rose water mist
Rose Water Soothing
salasylic acid
salicylic acid
% Salicylic acid
Salicylic acid
scent.Sweet Mint
sealWild rose
sesame seeds
shea balm

shea butter
Shea butter
Shea butter feel
shea butter herebecause
shea butter .I
Shea butter .Love
shea butter products
Shea butter soooo
shea moisture
Shea Moisture
sheer rose
sheer rose color
slight mint
slight mint tingle
soft rose
Soft Rose
Soft Rose Care
Soft Rose Lip
soft rose petals .This
soft rose scent

Les unités surlignées en jaune, sont celles qui sont considérées comme “parasites” et que le quatrième script supprime pour rendre des résultats “propres” dans les fichiers de sortie finaux.

L’extraction réalisée a donné en totalité 624 candidats termes en français et 438 en anglais (sans doublons) pour les différents nommages des ingrédients et actifs cosmétiques. En voici quelques uns (cf. listes complètes en annexe).

Extrait des candidats termes en français :

acide h
acide hialuronique
Acide hialuroniques
acide huyaluronique
acide hyalurinuque
acide hyaluronique
acide Hyaluronique
Acide hyaluronique
Acide Hyaluronique
acide hyaluroniques
Beure de Cacao
beurre de cacao
beurre de kariete
beurre de karite
beurre de karité
Beurre de karité
beurre de karité bio
beurre de karité brut

distillat de lavande
eau de bleuet
eau de bleuet
eau de Bleuet
eau de bleut
eau de lavande
eau de rose
Eau de Rose
eau de rose en lotion
eau essentielle de menthe
eau florale bio de lavande
eau florale bio d' hamamélis
eau florale de lavande
eau florale de rose
HE de lavande
HE de lavande fine
HE de Rose
huil d' argen
huile bio de jojoba
huile corporelle Rose
huile d' aloe
huile d' aloès
huile d' aloe vera
huile d' amande douce
huile d' amande douce
Huile d' amande douce
huile damandes
huile damandes douces
huiles végétales riche en Acides
huile végétal d' aléo
huile végétal d' aléo vera
huile végétal d' argan
huile vegetal de noisette
huile végétal de noisette
huile végétale bio noisette
huile végétale de jojoba
huile végétale de noisette
huile végétale de Noisette
huile végétale de rose
huile végétale de rose musquée
huile végétale jojoba
huille de rose
huille de rose musquée
HV d' argan

Extrait des candidats termes en anglais :

acid peels
aleo vera

aleo vera gel
aleo vers
almond oil
almond oils
aloe
aloe vera
aloe Vera
Aloe Vera
aloe vera gel
alpha hydroxy acids
argan oil
Argan oil
glycolic acid
Glycolic acid
glycolic acid mixture
Golden Jojoba
Golden Jojoba oil
hyaluronic acid
hydroxy acids
jajoba oil
jojoba esters
jojoba mix
jojoba oil
Jojoba oil
lavender oil
rose oil
rose water
Rose Water
rose water mist
Rose Water Soothing
shea butter
Shea butter
witch hazel
Witch hazel
Witch Hazel

4.4 Comparaison multilingue des résultats

Les résultats obtenus de l'extraction en français et en anglais ne sont pas égaux car les corpus analysés n'ont pas été constitués pour les mêmes besoins, ce qui justifie leur décalage en quantité de références à des ingrédients ou actifs cosmétiques.

Cela n'empêche pas d'avoir des résultats intéressants dans les deux langues et leur comparaison l'est tout autant.

En examinant les candidats termes des deux côtés, nous retrouvons les types de variations dont il est question dans l'état de l'art.

— Les variantes graphiques et orthographiques se distinguent bien dans les expressions relevées dans les deux langues. Le changement de casse est fréquent (indifféremment de l'emplacement de l'élément dans le terme) ainsi que la marque du pluriel :

Corpus français	Corpus anglais
acide **h**yaluronique	**r**ose **w**ater
acide **H**yaluronique	**R**ose **W**ater
Acide hyaluronique	**w**itch **h**azel
Acide **H**yaluronique	**W**itch **h**azel
acide hyaluroniqu**e**s	**W**itch **H**azel
	almond oil
	almond oil**s**

TABLE 4.1 – Variantes graphiques et orthographiques en français et anglais

Les fautes d'orthographe et de frappe sont également courantes ainsi que la manifestation ou non des caractères diacrités pour le français :

Corpus français	Corpus anglais
acide Gl**u**colique	j**a**joba oil
acide glyco**r**ique	al**eo** vera
huile **d**amandes douces	al**eo** ver**s**
huile **d**argan	Wit**x**h Hazel
beurre de kari**et**e	
beurre de karit**e**	
beurre de karit**é**	

TABLE 4.2 – Suite des variantes graphiques et orthographiques en français et anglais

— Les variations morphosyntaxiques qui se traduisent par l'insertion, la substitution ou la suppression d'un élément est très visible dans les résultats français ainsi que les ellipses dans les deux langues :

Corpus français	Corpus anglais
Huile de Rose	**the AHA**
huile de rose **de musquée**	**the aleo**
Huile de Rose **Musquée**	**The aloe**
huile de rose musquée **bio**	**the hazel**
huile de rose musquée **du chili**	**the hemp**
huile de rose **naturelle**	**the jojoba**
huile **végétale** de rose	**the mint**
huile **végétale** de rose **musquée**	**the Shea**
le bleuet	
le jojoba	
le karité	
les AHA	

TABLE 4.3 – Variantes morphosyntaxiques en français et anglais

— Les variantes sémantiques telles que les synonymes sont présentes dans les deux langues et les abréviations, que nous considérons comme des synonymes, le sont en français mais nous n'en trouvons pas en anglais :

Corpus français	Corpus anglais
distillat de lavande	**acid peels**
eau de lavande	**alpha hydroxy acids**
hydrolat de lavande sauvage	**hydroxy acids**
HA de rose	**Glycolic acid**
huile végétale de rose musquée	**fruit acids**
HV de rose musquée	
Huile Essentielle de Lavande	
HE de lavande fine	

TABLE 4.4 – Variantes sémantiques en français et anglais

Conclusion générale

L'extraction des différents nommages d'ingrédients et actifs cosmétiques à partir de corpus de commentaires réalisée dans ce travail, a obtenue de bons résultats pour le corpus français. Ils présentent de nombreuses expressions variées graphiquement et/ou morphosyntaxiquement pertinentes à intégrer dans les ressources linguistiques.

Le corpus anglais étant moins productif pour cette recherche, moins de patrons morphosyntaxiques ont été répertoriés, ce qui a réduit les chances d'extraire des candidats termes diversifiés. Malgré cela, quelques éléments intéressants se trouvent parmi les résultats anglais comme les synonymes de AHA.

Il s'avère que notre méthode d'extraction peut aller plus loin avec la découverte de nouveaux ingrédients si l'on introduit des éléments de base plus génériques comme "acid(e)" qui a permis d'identifier divers noms d'acides inattendus, sortant parfois du domaine de la cosmétique ("acide ascorbique", "acide citrique", "acide fucidine", "acide gamma-linolénique", "acide lactique", "acide laurique", "acide oméga", "acide pantothénique", "acide phosphorique", "acides aminés", "acides gras", "acide sacylique", "acide silicique minéral", "acide sté", "acide stéariqueel" en français et "Amino Acids", "cane acids", "cholic acid", "citric acid", "Fatty acids", "Kojic Acid", "lactic acid", "Salicylic acid" en anglais). Ils peuvent amener à faire des recherches pour se documenter dessus et être favorables à l'enrichissement de ressources linguistiques.

Il y a d'autres découvertes comme celle-ci, par exemple, deux variétés d'ingrédients de base ont été relevées grâce la distance de Levenshtein : "lavandin" et "menthol" qui peuvent aussi donner lieu à des recherches supplémentaires pour les ingrédients des produits cosmétiques.

Différents types d'ingrédients ont aussi été mis en évidence comme ceux de l'aloe vera ("gel Aloe Vera", "huile d' aloe vera", "jus d' aloe", "macerat d' aloes", "plante d' Aloe Vera", "pousse d' aloes"), ce qui peut également être une autre source d'étude.

Cette extraction propose donc un terrain propice aux découvertes malgré la présence de bruit parmi les candidats termes. En français, en raison d'un grand nombre de patrons morphosyntaxiques, des expressions extraites se chevauchent. Mais il s'agit d'un détail qu'il est possible de rectifier. Ce qui pourrait être intéressant à faire par la suite, serait d'entreprendre une extraction avec l'objectif inverse : prendre des types d'ingrédients comme éléments de base (huile, eau florale, beurre, gel, etc) pour analyser les ingrédients ramenés par cette méthode.

BIBLIOGRAPHIE

[Aubin and Hamon, 2006] Aubin, S. and Hamon, T. (2006). Improving term extraction with terminological resources. In *Advances in natural language processing*, Springer. – Cité page 17.

[Bourigault, 1994] Bourigault, D. (1994). *LEXTER, logiciel d'extraction de terminologie. Application à l'acquisition de connaissances à partir de textes*. PhD thesis, EHESS, Paris. – Cité page 17.

[Bourigault et al., 1996] Bourigault, D., Gonzalez-Mullier, I., and Gros, C. (1996). Lexter, a natural language processing tool for terminology extraction. *7th EURALEX International Congress*. – Cité pages 17, 18 et 20.

[Cabré et al., 2001] Cabré, M. T., Estopà, R., and Vivaldi, J. (2001). Automatic term detection : a review of current systems. In *Recent advances in computational terminology*, Amsterdam, Philadelphia. John Benjamins. – Cité page 17.

[Daille, 1994a] Daille, B. (1994a). *Approche mixte pour l'extraction automatique de terminologie : statistique lexicale et filtres linguistiques*. PhD thesis, Université Paris VII. – Cité page 17.

[Daille, 1994b] Daille, B. (1994b). Study and implementation of combined techniques for automatic extraction of terminology. In *The balancing act : Combining symbolic and statistical approaches to language, Workshop of the 32th annual meeting of the ACL*, Las Cruces, New Mexico, USA. – Cité pages 17, 18 et 20.

[Drouin and Langlais, 2006] Drouin, P. and Langlais, P. (2006). Evaluation du potentiel terminologique de candidats termes. In *Actes de JADT volume 2006*. – Cité page 18.

[Dunning, 1993] Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, 19 n1. – Cité page 20.

[Ibekwe-SanJuan, 2001] Ibekwe-SanJuan, F. (2001). *Extraction terminologique avec INTEX*. 4th annual INTEX workshop, Bordeaux. – Cité pages 17, 19 et 20.

[Jacquemin, 1997] Jacquemin, C. (1997). *Variation terminologique : reconnaissance et acquisition automatique de termes et de leurs variantes en corpus*. PhD thesis, IRIN, Nantes. – Cité pages 17 et 23.

[Korenchuk, 2014] Korenchuk, Y. (2014). *Extraction terminologique : vers la minimisation de ressources*. Brigitte Bigi. TALN-RECITAL, Marseille. Les actes de TALN-Recital 2014. – Cité pages 18, 19 et 20.

[Levenshtein, 1966] Levenshtein, V. I. (1966). Binary codes capables of correcting deletions, insertions and reversals. In *Sov. Phys. Dokl.*, volume 10, pages 707–710. – Cité page 25.

[L'Homme, 2004] L'Homme, M.-C. (2004). *La terminologie : principes et techniques*. Presses de l'Université de Montréal, Montréal. – Cité page 17.

[Morin and Daille, 2012] Morin, E. and Daille, B. (2012). Compositionnalité et contextes issus de corpus comparables pour la traduction terminologique. In *Conférence conjointe JEP-TALN-RECITAL 2012*. – Cité page 17.

[Rey, 1992] Rey, A. (1992). *La terminologie : noms et notions*. Collection Que sais-je ? Presses Universitaires de France, Paris, 2ème edition. – Cité page 17.

[Tartier, 2006] Tartier, A. (2006). Variation terminologique et analyse diachronique. In *Actes de la 13ème conférence sur le Traitement automatique des langues naturelles (TALN 2006)*, Leuven, Belgique. – Cité page 20.

[Vergne, 2003] Vergne, J. (2003). Un outil d'extraction terminologique endogène et multilingue. *Actes de TALN*, 2:139–148. – Cité page 19.

[Vergne, 2004] Vergne, J. (2004). Découverte locale des mots vides dans des corpus bruts de langues inconnues, sans aucune ressource. In *Actes de JADT*, Louvain, Belgique. – Cité page 19.

ANNEXE A

ANNEXE : SCRIPTS

A.1 Premier script

```
######################
ingred_base = ["chanvre", "cacao", "jojoba", "amande", "argan", "karité", "sésame", "noisette",
               "hamamélis", "bleuet", "oranger", "lavande", "menthe", "rose",
               "hyaluronique", "propolis", "aloe", "argile", "aha", "acide"]

# Ouverture du fichier étiqueté (sortie Treetagger)
fichier_ttg = open("ttg_corpus_commentaires_fr.txt", "r")
lignes = fichier_ttg.readlines()

# Création d'une liste 2D à partir du fichier ttg pour exploiter ses étiquettes
liste = []
for ligne in lignes :
    ligne = ligne.replace("\n","")
    liste.append(ligne)

liste2 = []
for element in liste :
    element = str(element)
    element = element.split("\t")
    liste2.append(element)

# Ouverture d'un fichier de sortie
fichier_variantes = open("variantes_levenshtein_corpus_commentaires_fr.csv", "w")

# Calcul des distances entre les éléments de base et les tokens
for ingred in ingred_base :
    for x in liste2 :
        distance = levenshtein_distance(ingred, x[0])
        print ("calculs")
        if distance <= 2 :
            fichier_variantes.write(ingred + "\t" + x[0] + "\t" + x[1] + "\t" + str(distance) + "\n")


fichier_ttg.close()
fichier_variantes.close()
```

FIGURE A.1 – Annexe : Premier script

Indication : Il s'agit de la suite du script après la fonction de la distance de Levenshtein.

A.2 Deuxième script

```
#!/usr/bin/python3.7
# -*- coding: <utf-8> -*-

ingred_base = ["chanvre", "cacao", "jojoba", "amande", "argan", "karité", "sésame", "noisette",
               "hamamélis", "bleuet", "oranger", "lavande", "menthe", "rose",
               "hyaluronique", "propolis", "aloe", "argile", "aha", "acide"]


fichier_variantes = open("variantes_levenshtein_integr_fr.txt", "r")
variantes = fichier_variantes.readlines()

# Creation d'une liste des variantes des éléments de base
ingred_variantes = []

for variante in variantes :
    variante = variante.strip()
    ingred_variantes.append(variante)

# Fusion des deux listes éléments de base + variantes
ingred_base.extend(ingred_variantes)
```

FIGURE A.2 – Annexe : Deuxième script (partie 1/3)

```
# Ouverture du fichier étiqueté (sortie Treetagger)
fichier_ttg = open("ttg_corpus_commentaires_fr.txt", "r")
lignes = fichier_ttg.readlines()

# Création d'une liste 2D à partir du fichier ttg pour exploiter ses étiquettes
liste = []
for ligne in lignes :
    ligne = ligne.replace("\n","")
    liste.append(ligne)

liste2 = []
for element in liste :
    element = str(element)
    element = element.split("\t")
    liste2.append(element)
```

FIGURE A.3 – Annexe : Deuxième script (partie 2/3)

```
# Ouverture d'un fichier de sortie avec les éléments matchés et leurs contextes
fichier_base = open("base_corpus_commentaires_fr.csv", "w")


# Stockage des index des éléments matchés
positions_base = []

for x in liste2:
    x.append(liste2.index(x))
    if x[0].lower() in ingred_base :
        positions_base.append(x[3])

# Extraction des contextes (fenêtre de 6 mots de chaque côté)
for p_base in positions_base :

    fichier_base.write(liste2[p_base -6][0] + "\t" + liste2[p_base -6][1] + "\t")
    fichier_base.write(liste2[p_base -5][0] + "\t" + liste2[p_base -5][1] + "\t")
    fichier_base.write(liste2[p_base -4][0] + "\t" + liste2[p_base -4][1] + "\t")
    fichier_base.write(liste2[p_base -3][0] + "\t" + liste2[p_base -3][1] + "\t")
    fichier_base.write(liste2[p_base -2][0] + "\t" + liste2[p_base -2][1] + "\t")
    fichier_base.write(liste2[p_base -1][0] + "\t" + liste2[p_base -1][1] + "\t")
    fichier_base.write(liste2[p_base][0] + "\t" + liste2[p_base][1] + "\t")
    fichier_base.write(liste2[p_base +1][0] + "\t" + liste2[p_base +1][1] + "\t")
    fichier_base.write(liste2[p_base +2][0] + "\t" + liste2[p_base +2][1] + "\t")
    fichier_base.write(liste2[p_base +3][0] + "\t" + liste2[p_base +3][1] + "\t")
    fichier_base.write(liste2[p_base +4][0] + "\t" + liste2[p_base +4][1] + "\t")
    fichier_base.write(liste2[p_base +5][0] + "\t" + liste2[p_base +5][1] + "\t")
    fichier_base.write(liste2[p_base +6][0] + "\t" + liste2[p_base +6][1] + "\t")
    fichier_base.write("\n")


fichier_variantes.close()
fichier_ttg.close()
fichier_base.close()
```

FIGURE A.4 – Annexe : Deuxième script (partie 3/3)

A.3 Troisième script

```
# Ouverture fichier de sortie avec les expressions matchées
fichier_pos = open("pos_corpus_commentaires_fr.txt", "w")

# Backslash utilisés pour revenir à la ligne pour plus de lisibilité
for p_base in positions_base :

    #monolexicaux
    if (liste2[p_base][1] == "NAM" or liste2[p_base][1] == "NOM") \
       and (liste2[p_base -1][1] == "DET:ART") :
        fichier_pos.write(liste2[p_base -1][0] + " " + liste2[p_base][0] + "\n")

    #polylexicaux : 2
    if (liste2[p_base][1] == "ADJ") \
       and (liste2[p_base +1][1] == "ADJ" or liste2[p_base +1][1] == "NOM" or \
            liste2[p_base +1][1] == "NAM") :
        fichier_pos.write(liste2[p_base][0] + " " + liste2[p_base +1][0] + "\n")

    if (liste2[p_base][1] == "ADJ") \
       and (liste2[p_base -1][1] == "ADJ"):
        fichier_pos.write(liste2[p_base -1][0] + " " + liste2[p_base][0] + "\n")

    if (liste2[p_base][1] == "NAM" or liste2[p_base][1] == "NOM") \
       and (liste2[p_base +1][1] == "ADJ" or liste2[p_base +1][1] == "ABR" or liste2[p_base +1][1] == "NAM" \
       or liste2[p_base +1][1] == "NOM") :
        fichier_pos.write(liste2[p_base][0] + " " + liste2[p_base +1][0] + "\n")

    if (liste2[p_base][1] == "NAM" or liste2[p_base][1] == "NOM") \
       and (liste2[p_base -1][1] == "NAM" or liste2[p_base -1][1] == "NOM"):
        fichier_pos.write(liste2[p_base -1][0] + " " + liste2[p_base][0] + "\n")
```

FIGURE A.5 – Annexe : Troisième script (partie 1/5)

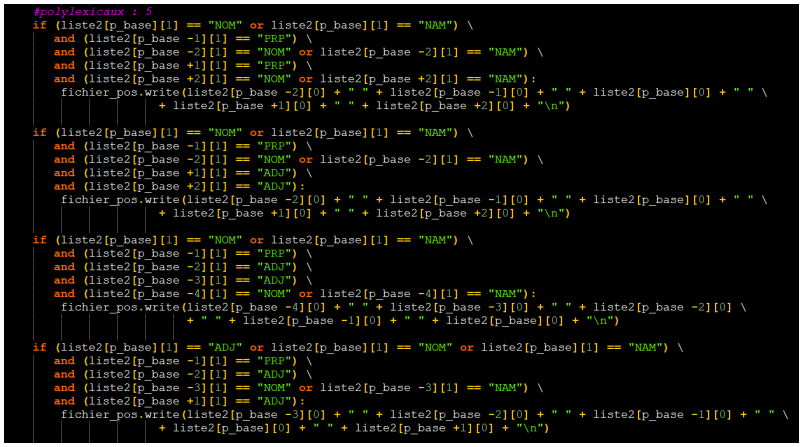
Indication : Le début de ce script est le même que dans le précédent.

```
    #polylexicaux : 3
    if (liste2[p_base][1] == "NOM" or liste2[p_base][1] == "NAM") \
       and (liste2[p_base -1][1] == "PRP") \
       and (liste2[p_base -2][1] == "ADJ" or liste2[p_base -2][1] == "NOM" or liste2[p_base -2][1] == "NAM"):
        fichier_pos.write(liste2[p_base -2][0] + " " + liste2[p_base -1][0] + " " + liste2[p_base][0] + "\n")

    if (liste2[p_base][1] == "NOM" or liste2[p_base][1] == "NAM") \
       and (liste2[p_base -1][1] == "ADJ") \
       and (liste2[p_base -2][1] == "NOM" or liste2[p_base -2][1] == "NAM"):
        fichier_pos.write(liste2[p_base -2][0] + " " + liste2[p_base -1][0] + " " + liste2[p_base][0] + "\n")

    if (liste2[p_base][1] == "NOM" or liste2[p_base][1] == "NAM") \
       and (liste2[p_base -1][1] == "NOM" or liste2[p_base -1][1] == "NAM") \
       and (liste2[p_base +1][1] == "NOM" or liste2[p_base +1][1] == "NAM" or liste2[p_base +1][1] == "ADJ"):
        fichier_pos.write(liste2[p_base -1][0] + " " + liste2[p_base][0] + " " + liste2[p_base +1][0] + "\n")

    if (liste2[p_base][1] == "NOM" or liste2[p_base][1] == "NAM") \
       and (liste2[p_base +1][1] == "ADJ") \
       and (liste2[p_base +2][1] == "ADJ"):
        fichier_pos.write(liste2[p_base][0] + " " + liste2[p_base +1][0] + " " + liste2[p_base +2][0] + "\n")

    if (liste2[p_base][1] == "ADJ") \
       and (liste2[p_base -1][1] == "PRP") \
       and (liste2[p_base -2][1] == "NOM" or liste2[p_base -2][1] == "NAM"):
        fichier_pos.write(liste2[p_base -2][0] + " " + liste2[p_base -1][0] + " " + liste2[p_base][0] + "\n")
```

FIGURE A.6 – Annexe : Troisième script (partie 2/5)

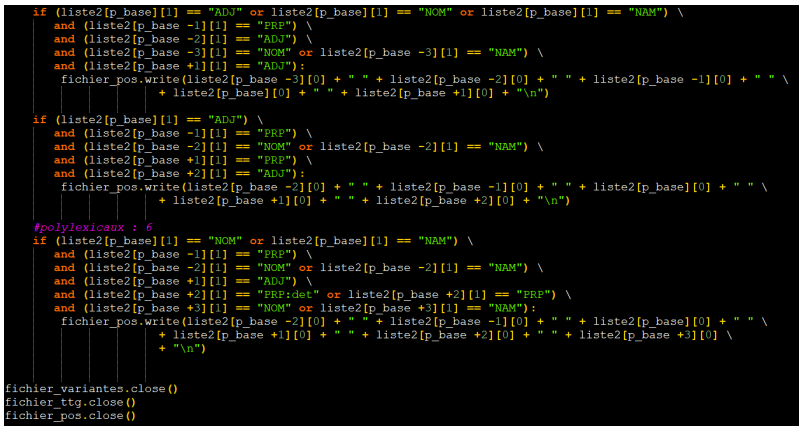
```
    #polylexicaux : 4
    if (liste2[p_base][1] == "NOM" or liste2[p_base][1] == "NAM") \
       and (liste2[p_base -1][1] == "PRP" or liste2[p_base -1][1] == "ADJ") \
       and (liste2[p_base -2][1] == "NOM" or liste2[p_base -2][1] == "NAM") \
       and (liste2[p_base +1][1] == "ADJ"):
        fichier_pos.write(liste2[p_base -2][0] + " " + liste2[p_base -1][0] + " " + liste2[p_base][0] + " " \
                          + liste2[p_base +1][0] + "\n")

    if (liste2[p_base][1] == "NOM" or liste2[p_base][1] == "NAM") \
       and (liste2[p_base -1][1] == "ADJ" or liste2[p_base -1][1] == "PRP") \
       and (liste2[p_base -2][1] == "ADJ") \
       and (liste2[p_base -3][1] == "NOM" or liste2[p_base -3][1] == "NAM"):
        fichier_pos.write(liste2[p_base -3][0] + " " + liste2[p_base -2][0] + " " + liste2[p_base -1][0] + " " \
                          + liste2[p_base][0] + "\n")

    if (liste2[p_base][1] == "NOM" or liste2[p_base][1] == "NAM") \
       and (liste2[p_base -1][1] == "PRP") \
       and (liste2[p_base -2][1] == "NOM" or liste2[p_base -2][1] == "NAM") \
       and (liste2[p_base +1][1] == "NOM" or liste2[p_base +1][1] == "NAM"):
        fichier_pos.write(liste2[p_base -2][0] + " " + liste2[p_base -1][0] + " " + liste2[p_base][0] + " " \
                          + liste2[p_base +1][0] + "\n")

    if (liste2[p_base][1] == "NOM" or liste2[p_base][1] == "NAM") \
       and (liste2[p_base -1][1] == "PRP") \
       and (liste2[p_base -2][1] == "NOM" or liste2[p_base -2][1] == "NAM") \
       and (liste2[p_base -3][1] == "NOM" or liste2[p_base -3][1] == "NAM"):
        fichier_pos.write(liste2[p_base -3][0] + " " + liste2[p_base -2][0] + " " + liste2[p_base -1][0] + " " \
                          + liste2[p_base][0] + "\n")

    if (liste2[p_base][1] == "ADJ") \
       and (liste2[p_base -1][1] == "PRP") \
       and (liste2[p_base -2][1] == "ADJ") \
       and (liste2[p_base -3][1] == "NOM" or liste2[p_base -3][1] == "NAM"):
        fichier_pos.write(liste2[p_base -3][0] + " " + liste2[p_base -2][0] + " " + liste2[p_base -1][0] + " " \
                          + liste2[p_base][0] + "\n")
```

FIGURE A.7 – Annexe : Troisième script (partie 3/5)



```
    #polylexicaux : 5
    if (liste2[p_base][1] == "NOM" or liste2[p_base][1] == "NAM") \
       and (liste2[p_base -1][1] == "PRP") \
       and (liste2[p_base -2][1] == "NOM" or liste2[p_base -2][1] == "NAM") \
       and (liste2[p_base +1][1] == "PRP") \
       and (liste2[p_base +2][1] == "NOM" or liste2[p_base +2][1] == "NAM"):
        fichier_pos.write(liste2[p_base -2][0] + " " + liste2[p_base -1][0] + " " + liste2[p_base][0] + " " \
                  + liste2[p_base +1][0] + " " + liste2[p_base +2][0] + "\n")

    if (liste2[p_base][1] == "NOM" or liste2[p_base][1] == "NAM") \
       and (liste2[p_base -1][1] == "PRP") \
       and (liste2[p_base -2][1] == "NOM" or liste2[p_base -2][1] == "NAM") \
       and (liste2[p_base +1][1] == "ADJ") \
       and (liste2[p_base +2][1] == "ADJ"):
        fichier_pos.write(liste2[p_base -2][0] + " " + liste2[p_base -1][0] + " " + liste2[p_base][0] + " " \
                  + liste2[p_base +1][0] + " " + liste2[p_base +2][0] + "\n")

    if (liste2[p_base][1] == "NOM" or liste2[p_base][1] == "NAM") \
       and (liste2[p_base -1][1] == "PRP") \
       and (liste2[p_base -2][1] == "ADJ") \
       and (liste2[p_base -3][1] == "ADJ") \
       and (liste2[p_base -4][1] == "NOM" or liste2[p_base -4][1] == "NAM"):
        fichier_pos.write(liste2[p_base -4][0] + " " + liste2[p_base -3][0] + " " + liste2[p_base -2][0] \
                  + " " + liste2[p_base -1][0] + " " + liste2[p_base][0] + "\n")

    if (liste2[p_base][1] == "ADJ" or liste2[p_base][1] == "NOM" or liste2[p_base][1] == "NAM") \
       and (liste2[p_base -1][1] == "PRP") \
       and (liste2[p_base -2][1] == "ADJ") \
       and (liste2[p_base -3][1] == "NOM" or liste2[p_base -3][1] == "NAM") \
       and (liste2[p_base +1][1] == "ADJ"):
        fichier_pos.write(liste2[p_base -3][0] + " " + liste2[p_base -2][0] + " " + liste2[p_base -1][0] + " " \
                  + liste2[p_base][0] + " " + liste2[p_base +1][0] + "\n")
```

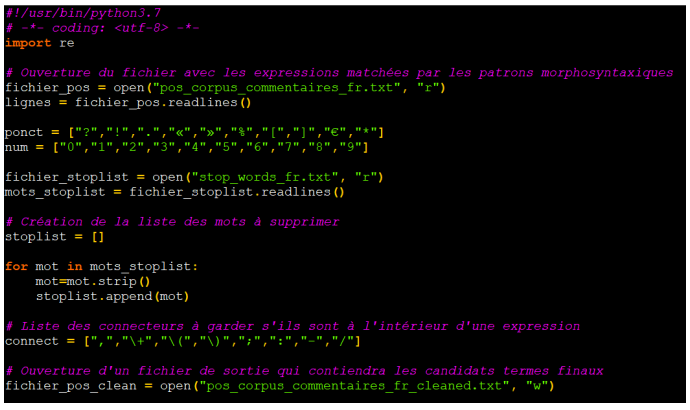
FIGURE A.8 – Annexe : Troisième script (partie 4/5)



```
    if (liste2[p_base][1] == "ADJ" or liste2[p_base][1] == "NOM" or liste2[p_base][1] == "NAM") \
       and (liste2[p_base -1][1] == "PRP") \
       and (liste2[p_base -2][1] == "ADJ") \
       and (liste2[p_base -3][1] == "NOM" or liste2[p_base -3][1] == "NAM") \
       and (liste2[p_base +1][1] == "ADJ"):
        fichier_pos.write(liste2[p_base -3][0] + " " + liste2[p_base -2][0] + " " + liste2[p_base -1][0] + " " \
                  + liste2[p_base][0] + " " + liste2[p_base +1][0] + "\n")

    if (liste2[p_base][1] == "ADJ") \
       and (liste2[p_base -1][1] == "PRP") \
       and (liste2[p_base -2][1] == "NOM" or liste2[p_base -2][1] == "NAM") \
       and (liste2[p_base +1][1] == "PRP") \
       and (liste2[p_base +2][1] == "ADJ"):
        fichier_pos.write(liste2[p_base -2][0] + " " + liste2[p_base -1][0] + " " + liste2[p_base][0] + " " \
                  + liste2[p_base +1][0] + " " + liste2[p_base +2][0] + "\n")

    #polylexicaux : 6
    if (liste2[p_base][1] == "NOM" or liste2[p_base][1] == "NAM") \
       and (liste2[p_base -1][1] == "PRP") \
       and (liste2[p_base -2][1] == "NOM" or liste2[p_base -2][1] == "NAM") \
       and (liste2[p_base +1][1] == "ADJ") \
       and (liste2[p_base +2][1] == "PRP:det" or liste2[p_base +2][1] == "PRP") \
       and (liste2[p_base +3][1] == "NOM" or liste2[p_base +3][1] == "NAM"):
        fichier_pos.write(liste2[p_base -2][0] + " " + liste2[p_base -1][0] + " " + liste2[p_base][0] + " " \
                  + liste2[p_base +1][0] + " " + liste2[p_base +2][0] + " " + liste2[p_base +3][0] \
                  + "\n")


fichier_variantes.close()
fichier_ttg.close()
fichier_pos.close()
```

FIGURE A.9 – Annexe : Troisième script (partie 5/5)

A.4 Quatrième script



```
#!/usr/bin/python3.7
# -*- coding: <utf-8> -*-
import re

# Ouverture du fichier avec les expressions matchées par les patrons morphosyntaxiques
fichier_pos = open("pos_corpus_commentaires_fr.txt", "r")
lignes = fichier_pos.readlines()

ponct = ["?","!",".","«","»","$","[","]","€","*"]
num = ["0","1","2","3","4","5","6","7","8","9"]

fichier_stoplist = open("stop_words_fr.txt", "r")
mots_stoplist = fichier_stoplist.readlines()

# Création de la liste des mots à supprimer
stoplist = []

for mot in mots_stoplist:
    mot=mot.strip()
    stoplist.append(mot)

# Liste des connecteurs à garder s'ils sont à l'intérieur d'une expression
connect = ["'","\+","\(","\)",";",":","-","/"]

# Ouverture d'un fichier de sortie qui contiendra les candidats termes finaux
fichier_pos_clean = open("pos_corpus_commentaires_fr_cleaned.txt", "w")
```

FIGURE A.10 – Annexe : Quatrième script (partie 1/3)

```
for ligne in lignes :
    # suppression des signes de ponctuations
    for p in ponct :
        ligne = ligne.replace(p, " ")
        ligne = ligne.strip()

    # suppression des chiffres
    for n in num :
        ligne = ligne.replace(n, "")
        ligne = ligne.strip()
        ligne = ligne + "\n"

    # uniformisation des espaces (1 espace entre chaque mot)
    if " "+" "+" " in ligne :
        ligne = ligne.replace(" "+" "+" ", " ")

    if " "+" " in ligne :
        ligne = ligne.replace(" "+" ", " ")

    # suppression des connecteurs s'ils sont à l'extérieur d'une expression
    for c in connect :

        ##cas n°1 : connecteur à l'extérieur en début de ligne
        if re.search("^" + c, ligne, flags=re.I) :
            ligne = re.sub("^" + c, "", ligne, flags=re.I)

        ##cas n°2 : connecteur à l'extérieur en fin de ligne
        if re.search(c + "$", ligne, flags=re.I) :
            ligne = re.sub(c + "$", "", ligne, flags=re.I)
```

FIGURE A.11 – Annexe : Quatrième script (partie 2/3)

```
    # suppression des mots de la stoplist
    for m in stoplist :

        ##cas n°1 : mot entre deux espaces
        if re.search(" " + m + " ", ligne, flags=re.I) :
            ligne = re.sub(" " + m + " ", " ", ligne, flags=re.I)

        ##cas n°2 : mot en début de ligne
        if re.search("^" + m + " ", ligne, flags=re.I) :
            ligne = re.sub("^" + m + " ", "", ligne, flags=re.I)

        ##cas n°3 : mot outil en fin de ligne
        if re.search(" " + m + "$", ligne, flags=re.I) :
            ligne = re.sub(" " + m + "$", "", ligne, flags=re.I)

    #suppression des connecteurs collés à un mot de la stoplist
    for c in connect :
        for m in stoplist :
            if re.search(c + m, ligne, flags=re.I) :
                ligne = re.sub(c + m, "", ligne, flags=re.I)


    fichier_pos_clean.write(ligne)


fichier_pos.close()
fichier_pos_clean.close()
fichier_stoplist.close()
```

FIGURE A.12 – Annexe : Quatrième script (partie 3/3)

ANNEXE B

ANNEXE : LISTES DES CANDIDATS TERMES EXTRAITS

B.1 Candidats termes finaux en français

A(a base d' acides
acide ascorbique
acide citrique
acide fucidine
acide gamma-linolénique
acide Glucolique
acide glycolique
Acide Glycolique
acide glycorique
acide h
acide hialuronique
Acide hialuroniques
acide huyaluronique
acide hyalurinuque
acide hyaluronique
acide Hyaluronique
Acide hyaluronique
Acide Hyaluronique
Acide Hyaluronique
acide hyaluroniques
acide Illoronique
acide lactique
acide laurique
acide oméga
acide pantothénique
acide phosphorique
acide sacylique
acide salicilique
acide salicylique
acide salycilique
acides aminés
acides gras
Acides Gras

acides gras essentiels
acides gras insaturés
acide silicique
acide silicique minéral
acide sté
acide stéariqueel
acid oily
actif AHA
âge gel d' aloé
aha
AHA
AHA argile
AHA argile verte
AHA naturel
alcola de lavande
aléo vera
aloe
Aloe
aloé
ALOE ACTIVATOR
aloé buvable
aloé degonfle
aloe evera
aloe http :// aroma-zone com/arom
aloe lotion
Aloe propolis
aloe pure
aloes vera
aloès vera
aloe vera
aloe Vera
Aloe vera
Aloe Vera
ALOE Vera
ALOE VERA
aloé vera
aloé Vera
Aloé vera
Aloé Vera
aloe véra
aloe Véra
Aloe Véra
aloé véra
Aloé véra
Aloé Véra
aloé vera frais
Aloe Vera/huile
Aloe Vera/huile de chanvre
aloe vera naturelle

aloe vera pur
aloé vera pur
aloe vera pure
aloé vera réparatrice
Aloé Veré
amande amère
amande douce
amandes grillées
ammande
argan
argan cosmétique
argane
argane cosmétique
argane pure
argan naturel
argan pure
argen
argile
argile algues
argile blanche
ARGILE CATTIER
argile écocert
argile ionique
argile jaune
argile Juvaflorine
argile rose
argile rouge
argiles
argile sebo
argile sèche
argiles écocert
argiles vertes
argile verte
Argile verte
argile vertelaçonreme
argile vertelaçonreme refine
argil verte
avance beurre de karite
bardane eau de rose
base d' acide
base d' acide Glucolique
base d' acide glycolique
base d' acide hyaluronique
base d' acides
base d' acide salicylique
base d' acides de fruits
base d' AHA
base d' AHA
base d' aloe

base d’ aloé
base d’ Aloé
base d’ aloès
base d’ aloe vera
base d’ aloé vera
base d’ Aloé Vera
base d’ argan
base d’ argile
base d’ argile
base d’ argile blanche
base d’ argile rose
base d’ argile verte
base de karité
base de menthe
bcp de cacao
bcp de cacao en poudre
besoin d’ acide
besoin d’ acide Hyaluronique
Beure de Cacao
beurre de cacao
beurre de kariete
beurre de karite
beurre de karité
Beurre de karité
beurre de karité bio
beurre de karité brut
bio de jojoba
bio de lavande
BIO de lavande
bio d’ hamamélis
Bleu d’ Argan
bleuet
boreade : argile
boreade : argile verte
bourrache de sésame
boutons de rose
Boutons de rose
boutons de roses
brume de lavande
buvable d’ aloe
cataplasmes d’ argile
champs de lavande
chantilly de Karité
chanvre
chanvre pure
cire de jojoba
Citric acid
comparaison d’ amande
concentration d’ acide

concentration d' acide de fruit
concentration en Aha
couche d' argile
cure naturelle d' hamamelis
d' acides
d' AHA
d' aloe
d' aloé
daloe vera
d' aloé vera
d' amande
d' amande douce
damandes douces
dargan
d' argile
d' argile blanche
de jojoba
de lavande
de lavande vraie
de rose
de Rose
de rose de puressentiel
de rose en HE
distillat de lavande
eau de bleuet
eau de bleuet
eau de Bleuet
eau de bleut
eau de lavande
eau de rose
Eau de Rose
eau de rose en lotion
eau essentielle de menthe
eau florale bio de lavande
eau florale bio d' hamamélis
eau florale de lavande
eau florale de rose
E Huile d' argan
émulsion d' aloès
épine de rose
essence de lavande
essenssielle de lavande
essentiel de lavande
essentielle de lavande
essentielle de Lavande
Essentielle de Lavande
essentielle de menthe
essentielle rose
essentielles de lavande

feuilles d’ aloès
feuilles de menthe
feuilles de menthe fraiches
feuilles de menthe poivrée
feuilles fraîches d’ Aloé
feuilles fraîches d’ Aloé vera
fleur d’ oranger
fleur d’ oranger
fleurs de lavande
fleurs de lavande fraîches
Fleurs d’ oranger
florale de lavande
fois argile
fois argile sèche
fraîches d’ Aloé
gamme d’ aloe
gamme en aloe
gamme en aloe vera
gel Aloe
Gel Aloe
gel aloé
gel Aloé
gel aloes
gel Aloe Vera
Gel Aloe vera
gel aloé vera
gel Aloé vera
gel aloé véra
gel Aloé Veré
gel d’ aloe
gel d’ Aloe
gel d’ aloé
gel d’ aloé
gel d’ Aloé
gel d’ aloé buvable
gel d’ aloes
gel d’ aloès
gel d’ aloès vera
gel d’ aloe vera
gel d’ Aloe vera
gel d’ Aloe Vera
gel d’ aloé vera
gel d’ aloé véra
gel d’ aloé vera en pharmacie
gel d’ aloe vera pur
gel d’ aloé vera pur
gel en aloès
gels d’ aloe
Gommage amande

goutte de lavande
goutte de lavande fine
goutte de noisette
goutte de noisette sur peau
graines de sésame
guide Aloé
guide Aloé Véra
HA de lavande
HA de menthe
HA de rose
hamamelis emarronier
hamamelis virginia
HE de lavande
HE de lavande fine
HE de Rose
huil d' argen
huile bio de jojoba
huile corporelle Rose
huile d' aloe
huile d' aloès
huile d' aloe vera
huile d' amande
huile d' amande
Huile d' amande
huile d' amande douce
huile d' amande douce
Huile d' amande douce
huile damandes
huile damandes douces
huile dargan
huile d' argan
huile d' argan
huile d' Argan
Huile d' argan
huile d' argan cosmétique
huile d' argane
huile d' argane cosmétique
huile d' argane pure
huile d' argan naturel
huile d' argan pure
huile d' argan sur visage
huile d' argon
huile de chanvre
huile de chanvre pure
huile de jojba
huile de jojoba
huile de jojoba
huile de Jojoba
Huile de Jojoba

huile de jojoba Suisse
huile de karité
huile de lavande
Huile de Lavande
huile de noisette
huile de noisette e
Huile de noisettes
huile de rose
Huile de Rose
huile de rose de musquée
huile de rose musquée
Huile de Rose Musquée
huile de rose musquée bio
huile de rose musquée du chili
huile de rose naturelle
huile essenssielle de lavande
huile essentiel de lavande
huile essentiel de lavande vrai
huile essentielle de lavande
huile essentielle de Lavande
Huile Essentielle de Lavande
huile essentielle de lavande vrai
huile essentielle de lavande vraie
huile essentielle de rose
huile essentielles de lavande
huile jojoba
Huile jojoba
HUILE NOISETTE
HUILE ROSE
HUILE ROSE MUSQUEE
huiles d' argan
huiles de noisette
huiles de rose
huiles de rose musque
huiles végétales riche en Acides
huile végétal d' aléo
huile végétal d' aléo vera
huile végétal d' argan
huile vegetal de noisette
huile végétal de noisette
huile végétale bio noisette
huile végétale de jojoba
huile végétale de noisette
huile végétale de Noisette
huile végétale de rose
huile végétale de rose musquée
huile végétale jojoba
huille de rose
huille de rose musquée

HV d' argan
HV de jojoba
HV de jojoba contre
HV de noisette
HV de rose
HV de rose musquée
HV jojoba
HV Jojoba
HV rose
HV Rose
HV rose musqué
HV rose musquée
HV Rose Musquée
hyalurinuque Lefery
hyaluronique
Hyaluronique
hyaluroniques morphine
hydrolat de lavande
hydrolat de lavande sauvage
hydrolat de menthe
hydrolat de menthe poivrée
hydrolat de rose
Hydrolat de rose
Hydrolat de rose de Damas
hydrolat d' hamamélis
Hydrolat d' hamamélis
infusion de menthe
infusion menthe
injections acide
injections d' acide
injections d' acide hyaluronique
injections d' acide hyaluroniques
Jojba
jojoba
jojoba
jojoba de rose
jojoba Suisse
jus buvable d' aloe
jus buvable d' aloe vera
jus d aloe
jus d' aloe
jus d' aloe
jus d' aloé
jus d' aloe pure
jus d' aloes
jus d' aloès
jus d' aoloes
karité
karite bio

karité bio
karité brut
karite marker
l’ acide
l’ Acide
L’ acide
lait d’ amande
la jojoba
la lavande
la Lavande
La lavande
l’ aloe
l’ Aloe
L’ aloe
L’ Aloe
L’ ALOE
l’ aloé
l’ Aloé
L’ aloé
l’ aloes
l’ aloès
L’ aloès
L’ Aloès
l’ amande
la menthe
la Menthe
La menthe
la propolis
la Propolis
La propolis
l’ argan
l’ argil
largile
l’ argile
L’ argile
largile verte
la rose
lavande
Lavande
lavande aspic
lavande fine
LAVANDE FINE
lavande fraîches
lavande officinale
lavande rafraîchisssante
lavande sauvage
lavande vrai
lavande vraie
lavandin

le AHA
le bleuet
le jojoba
le karité
le lavandin
le menthol
le rose
les acides
les Acides
les AHA
Les Aha
Les AHA
les amandes
l'essence de lavande
l' hamamélis
l'HV de Jojoba
L'hydrolat de menthe
lotion menthe
lotion menthe poivrée
luile d' ammande
macerat d' aloes
marque argile
marque Bleu d' Argan
marsques d' argile
Masque argile
masque d' argile
masque d' argile
masque d' argile rose
masque d' argyle
masques(argile
masques argiles
masques argiles écocert
masques(argile verte
masques d' argile
masques d' argile
masques d' argiles
masques d' argiles
masques d' argiles vertes
masques d' argile verte
mélange d' argiles
mélange entre cacao
menthe
menthe antiseptique
menthe fraiches
menthe poivrée
menthe poivrée http ://bellaunaturel over-blog
com/article-lotion-astringente-au-persil-a-la-menthe-poivree- html
menthe verte
Menthe verte

mesmasques d' argiles
micro noisette
moment huile de jojoba
mondial d' aloès
naturelle d' hamamelis
neostrata aha
noisette e
odeur de lavande
odeur de rose
oranger
pacs d' argile
part argan
pate d' argile
pierre d' argile
plante d' aloe
plante d' Aloe
plante d' aloès
plante d' aloès en nature
plante d' aloe vera
plante d' Aloe Vera
plante d' aloe vera en floriculture
plante verte d' aloe
plante verte d' aloe vera
poudre d' amande
poudre d' amandes
Pour aloé
pourcentage d' aloe
pousse d' aloes
présente d' aloe
présente d' aloe vera
propolis
Propolis
recherche Aloe
recherche Aloe vera
riche en Acides
Riche en acides
rose
rose claire
rose clairs
rose -hydratation+hydratation
rose Marjorie
rose musque
rose musqué
rose musquee
ROSE MUSQUEE
rose musquée
Rose musquée
Rose Musquée
rose musquée bio

rose naturelle
rose rouge
rose sublime
savon d' argile
savon d' argile rouge
scientifique pour argen
semaine argile
semaine argile verte
sève aloé
sève aloé Vera
soins aha
Solaire Aloès
sortes d' Aloe
sortes d' Aloès
sortes d' Aloe Vera
soupe d' argile
soupe d' argile blanche
soupe d' argile verte
soupe d' argile verte en poudre
SweatStop® Aloe
SweatStop® Aloe Vera
terre d' argile
terre d' argile verte
test avec jojoba
tube d' argile
tube d' argile rose
type argile
type d' aloe
type huile d' argan
type jojoba
type lavande
type lavande officinale
un acide
un Aloé
une argile
une noisette
végétal d' aléo
végétal d' argan
vegetal de noisette
végétal de noisette
végétale de jojoba
végétale de noisette
végétale de Noisette
végétale de rose
Vera/huile de chanvre
verte d' aloe
vie en rose
vinaigre de lavande
vit A acide

vitamine A acide
Vitamine A acide
VITAMINE A ACIDE
vitamine A acide en traitement

B.2 Candidats termes finaux en anglais

acid
acid absorption
acid lotion
acid mixture
acid peels
acids
acid wash
aha
aha lotion
AHA soap
AHA toner
a lavender
Alcohol-Free Rose
Alcohol-Free Rose Petal
aleo vera
aleo vera gel
aleo vers
almond
almond meal
Almond meal
almond oil
almond oils
aloe
Aloe lip
Aloe lip tin
aloe tin
Aloe Tin
Aloe TinCarmen
Aloe TinGreat
Aloe TinGreat lip
Aloe TinThis
aloe Vaseline
aloe vera
aloe Vera
Aloe Vera
aloe vera formula
aloe vera gel
Aloe Vera type
Aloe Vera version
aloe version
alotOriginal rose

alpha hydroxy acids
Amino Acids
a mint
Andes mint
argan
argan oil
Argan oil
a rose
artificial rose
a shea
attractive rose
attractive rose scent
awesome mint
awesome mint flavorI
beautiful rose
beautiful rose tint
bizarre mint
bizarre mint vanilla
Blueberry Rose
Blueberry Rose Salve
cacao
cacao con
cacao de
cacao nunca
cane acids
chocolate mint
chocolate orange
cholic acid
citric acid
classic roses
comon orange
cool mint
Cool mint
cool mint flavor
Cool mint kind
crisp mint
dark rose
dark rose color
de cacao
deep orange
de Rose
De Rose
de Rose Nutriuleur
Dior rose
distinct rose
distinct rose smell
Enchanted Rose
Enchanted Rose swatch
enouigh mint

enouigh mint scent
Eucalyptus Mint
everything mint
excellent mint
excellent mint feelI
Fatty Acid
Fatty acids
fresh mint
fresh mint cube
fresh mint flavor
fresh mint scent
fresh mint smell
fresh rose
fresh rose balm
fresh roses
fruit acids
gentle rose
gentle rose scent
glycolic acid
Glycolic acid
glycolic acid mixture
Golden Jojoba
Golden Jojoba oil
good mint
goodThe mint
great mint
great mint smell
green mint
green orange
Green Orange
Hazel Alcohol
Hazel Alcohol Free
hazel eyes
Hazel Facial
Hazel Facial TonerI
Hazel products
hazel rose
hazel rose petal
hazel toner
Hazel Toner
Hazel Tonergreat
HEMP
hemp body
hemp body shop
HEMP products
hyaluronic acid
hydroxy acids
instantlyThe rose
instantlyThe rose color

it Cucumber mint
jajoba oil
jojoba esters
jojoba mix
jojoba oil
Jojoba oil
kojic acid
Kojic Acid
kojic acid lotion
lactic acid
lavender hue
lavender mint
lavender oil
light rose
light rose color
light rose scent
light rose tint
like roses
Loves roses
main acid
mandarin green orange
manHyaluronic acid
mask shea
mild mint
mild rose
mild rose smell
mint
Mint
mint addition
Mint balm
mint burts
mint burts bees
mint chap
mint Chap
mint chapstick
mint chap stick
mint Chap Stick
mint chapsticks
mint chocolate
mint Coco
mint cocoa
mint cocoa lip
mint condition
mint cooling
mint cube
Minted Rose
Minted Rose Lip
mint eos
mint exfoliator

mint feel
mint feelI
mint feeling
mint flavor
Mint flavor
MINT FLAVOR
mint flavor ChapStick
mint flavor eos
mint flavorI
mint flavoring
mint flavor lip
mint flavour
Mint Jack
Mint Jack Black
Mint Juleps
mint Julio
Mint Julips
Mint Julips balm
mint kind
mint lip
mint lip balms
mint lip exfoliator
mint lip products
mint maniac
Mint Maniac
mint maniac exfoliator
mint maniac variety
mint oneThere
mint products
mint scent
Mint scent/flavor
mint scrub
mint shower
mint shower gel
mint smell
mint taste
mint tingle
mint type
mint vanilla
mint vanilla flavor
Mint variety
mint version
Mocha Rose
month rose
Morbid Rose
Moroccan clay
Nivea Rose
normal mint
occitane shea

occitane shea butter
oil/ jojoba
Orange
ORANGE
orange callipo
orange cleanser
orange color
orange colour
orange essence
Orange Essence
Orange Essence cleanser
orange flavor
orange mango
orange neon
orange scent
Orange shade
orange tint
orange versions
raspberry rose
Raspberry Rose
raspberry rose kiss
Raspberry Rose KissClassic
Raspberry Rose KissLove
Raspberry Rose KissLoveeeeee
Raspberry Rose KissThe
Red Rose
Red Rose Tonique
refreshedLove shea
refreshedLove shea butter
regular mint
regular shea
regular shea butter
rose
Rose
rose balm
rose bud
rosebud rose
rose bud salve
Rose Care
Rose Care Balm
rose color
Rose color
rose colour
rose colour tint
rose flavour
rose gold
Rose hip
Rose Hip
Rose Hip Oil

Rose hip serum
rose kiss
Rose KissClassic
rose kiss lip
Rose KissLove
Rose KissLoveeeeee
Rose KissThe
Rose KissThe Milani
Rose Lip
Rose Lip Balm
Rose Lip BalmIt
Rose Lip BalmLove
Rose Lip BalmSoft
Rose Nutriuleur
rose oil
rose petal
Rose Petal
rose petals
Rose Petal Witch
rose products
Rose salve
Rose Salve
rose scent
Rose scent
rose scents
rose sent
rose sent/taste
rose shade
Rose Shine
Rose Shine Lip
rose smell
Rose smell
rose smellLove
Rose swatch
rose tint
Rose Tonique
Rose treatment
Rose varieties
rose water
Rose Water
rose water mist
Rose Water Soothing
salasylic acid
salicylic acid
Salicylic acid
scent Sweet Mint
sealWild rose
sesame seeds
shea balm

shea butter
Shea butter
Shea butter feel
shea butter products
shea moisture
Shea Moisture
sheer rose
sheer rose color
slight mint
slight mint tingle
soft rose
Soft Rose
Soft Rose Care
Soft Rose Lip
soft rose petals
soft rose scent
spear mint
spear mint flavorI
strong mint
Strong mint
Strong mint scent
strong mint smell
subtle mint
subtle mint taste
subtle rose
subtle rose scent
sugar mint
sugar mint scrub
sugar rose
Sugar Rose
Sugar Rose Shine
Sugar Rose varieties
super orange
sweet mint
Sweet mint
Sweet Mint
sweet mint eos
sweet mint flavor
sweet mint scent
sweet mint taste
sweet rose
sweet rose scent
taste The mint
tasty mint
tasty mint flavor
the acids
the AHA
the aleo
the aloe

the Aloe
The aloe
the hazel
the hemp
the jojoba
the mint
the Mint
The mint
THE MINT
The mint chap
The mint cube
The mint flavor
THE MINT FLAVOR
the orange
the Orange
The orange
THE ORANGE
Therapy® Aloe
Therapy® Aloe Tin
Therapy® Aloe TinCarmen
Therapy® Aloe TinGreat
Therapy® Aloe TinThis
the rose
the Rose
The rose
The Rose
The Rose Water
the shea
the Shea
The Shea
The Shea butter
tingly mint
tingly mint flavor
toner/witch hazel
vanilla mint
Vanilla Mint
Vanilla Mint balm
vanilla mint flavor
vanilla mint scent
Vaseline aloe
Vaseline Aloe
Vaseline Aloe lip
Vaseline aloe tin
witch hazel
Witch hazel
Witch Hazel
Witch Hazel Alcohol
Witch Hazel Facial
witch hazel rose

witch hazel toner
Witch Hazel Toner
Witch Hazel Tonergreat
Witxh Hazel
Witxh Hazel products