Apprentissage de connaissances morphologiques pour l'acquisition automatique de ressources lexicales

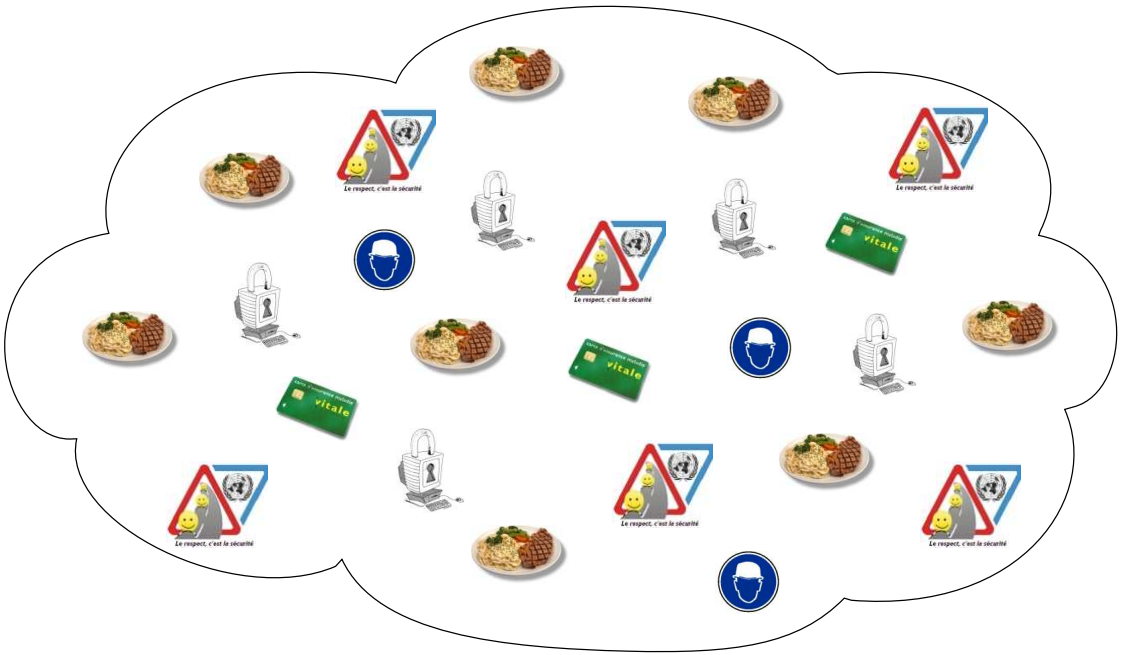
Delphine Bernhard

Laboratoire TIMC-IMAG, Grenoble

30 novembre 2006

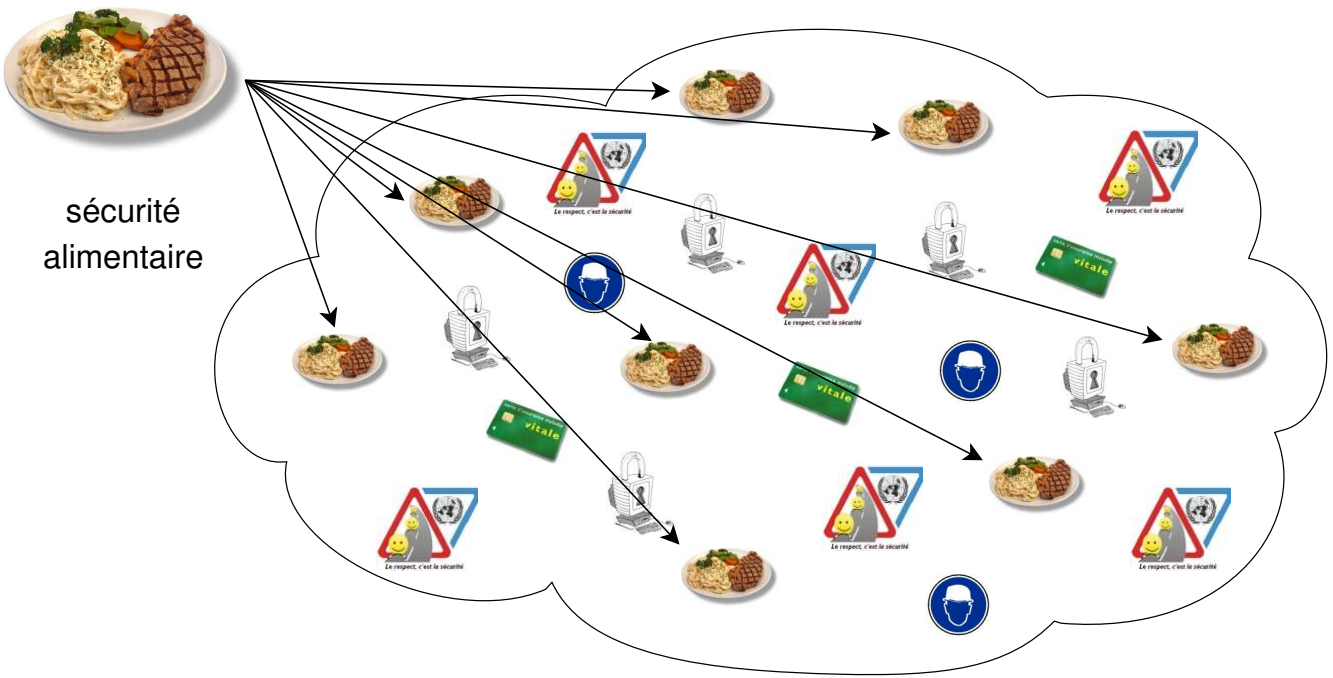
Contexte : organisation des données

Sécurité ?



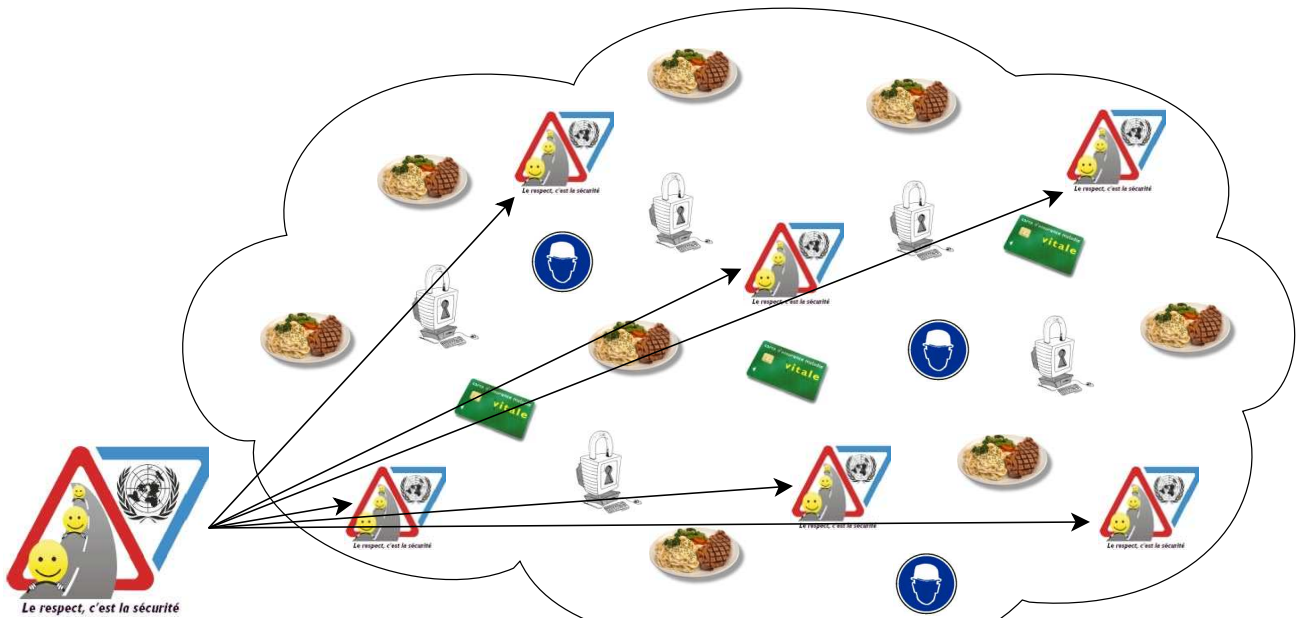
Contexte : organisation des données

Sécurité ?



Contexte : organisation des données

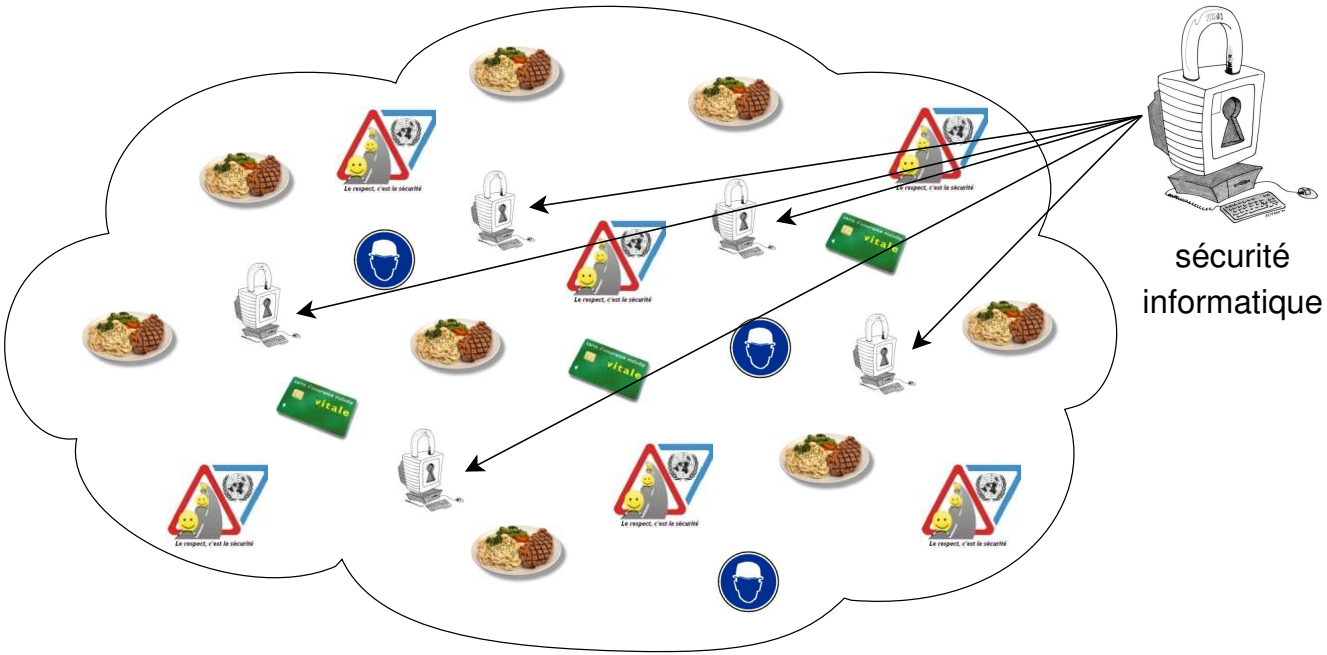
Sécurité ?



sécurité routière

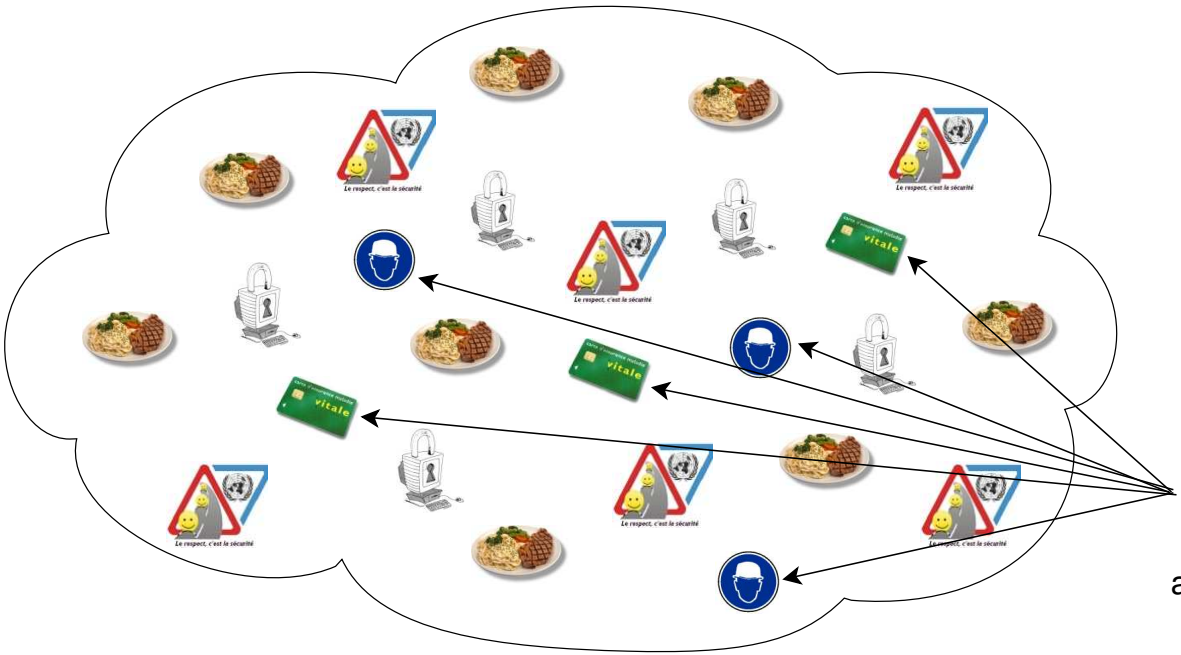
Contexte : organisation des données

Sécurité ?



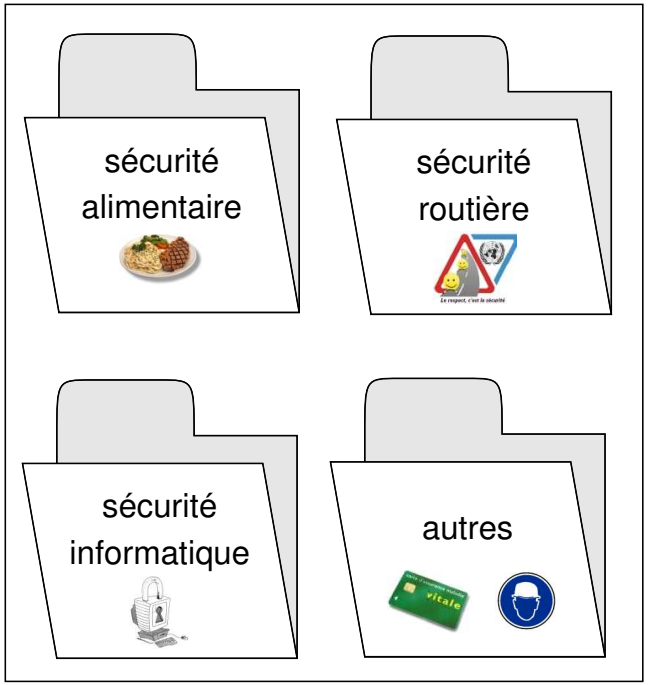
Contexte : organisation des données

Sécurité ?



Contexte : organisation des données

Sécurité ?



Organisation et structuration des informations via des ressources décrivant et classant les connaissances

Quelles ressources pour décrire les connaissances ?

Ressources construites manuellement par des experts

dictionnaires, terminologies, thésaurus, ontologies

\+ contrôle, précision

– coût, couverture

Descriptions collaboratives

indexation sociale (folksonomie) : tags $\sim$ mots-clés

\+ faible coût, couverture

– imprécision, absence de contrôle

Construction automatique de ressources

- Identification des termes représentant les concepts
- Acquisition de relations sémantiques pour la structuration des connaissances

Acquisition automatique de ressources lexicales

Patrons

- Termes complexes
  N + Adj
  éruption vulcanienne
- Relations sémantiques
  SN tels que SN+ ou SN
  des phénomènes climatiques **tels que** la température ou les précipitations

Statistiques

- Termes simples
  Mesures de comparaison
  Log du rapport de vraisemblance
- Termes complexes
  Mesures d'association
  Information mutuelle, χ^2, coefficient de Jaccard, etc.
- Similarité contextuelle
  Comparaison de vecteurs de co-occurrence

Rôle de la morphologie

Morphologie

Étude des morphèmes = les plus petites unités linguistiques qui allient forme et sens

Rôle de la morphologie

Morphologie

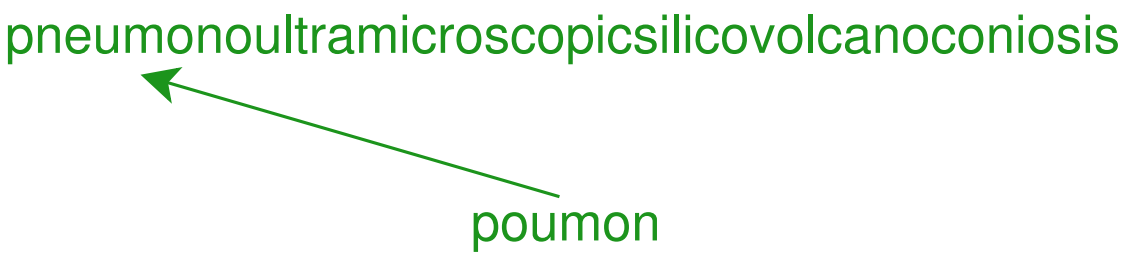
Étude des morphèmes = les plus petites unités linguistiques qui allient forme et sens

pneumonoultramicroscopicsilicovolcanoconiosis

Rôle de la morphologie

Morphologie

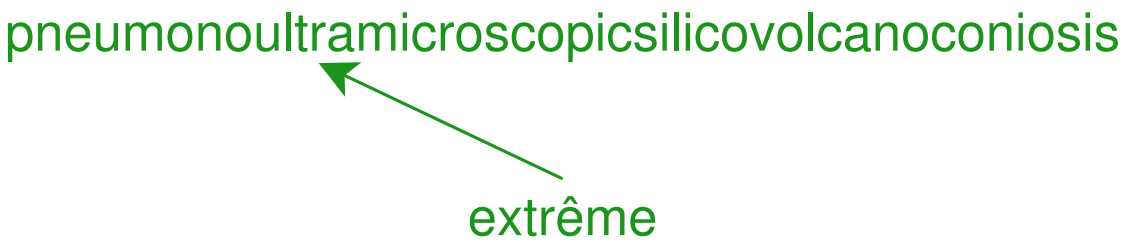
Étude des morphèmes = les plus petites unités linguistiques qui allient forme et sens



Rôle de la morphologie

Morphologie

Étude des morphèmes = les plus petites unités linguistiques qui allient forme et sens



Rôle de la morphologie

Morphologie

Étude des morphèmes = les plus petites unités linguistiques qui allient forme et sens

pneumonoultramicroscopicsilicovolcanoconiosis

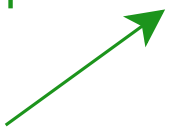
microscopique

Rôle de la morphologie

Morphologie

Étude des morphèmes = les plus petites unités linguistiques qui allient forme et sens

pneumonoultramicroscopicsilicovolcanoconiosis

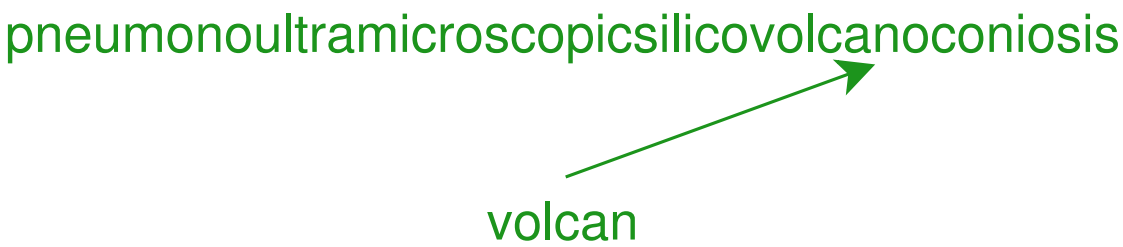


silicium

Rôle de la morphologie

Morphologie

Étude des morphèmes = les plus petites unités linguistiques qui allient forme et sens



Rôle de la morphologie

Morphologie

Étude des morphèmes = les plus petites unités linguistiques qui allient forme et sens

pneumonoultramicroscopicsilicovolcanoconiosis

poussière

Rôle de la morphologie

Morphologie

Étude des morphèmes = les plus petites unités linguistiques qui allient forme et sens

pneumonoultramicroscopicsilicovolcanoconiosis

atteinte

Rôle de la morphologie

Morphologie

Étude des morphèmes = les plus petites unités linguistiques qui allient forme et sens

pneumonoultramicroscopicsilicovolcanoconiosis

maladie des poumons résultant de l'inhalation de poussières de silicium très fines produites par des volcans

Rôle de la morphologie

Morphologie

Étude des morphèmes = les plus petites unités linguistiques qui allient forme et sens

pneumonoultramicroscopicsilicovolcanoconiosis

maladie des poumons résultant de l'inhalation de
poussières de silicium très fines produites par des volcans

Identification de termes

Vocabulaire technique : utilisation fréquente de morphèmes caractéristiques comme méga+, micro+, +gramme ou +graphe.

Rôle de la morphologie

Morphologie

Étude des morphèmes = les plus petites unités linguistiques qui allient forme et sens

pneumonoultramicroscopicsilicovolcanoconiosis

maladie des poumons résultant de l'inhalation de
poussières de silicium très fines produites par des volcans

Identification de termes

Vocabulaire technique : utilisation fréquente de morphèmes caractéristiques comme méga+, micro+, +gramme ou +graphe.

Extraction de relations sémantiques

cobalto**thérapie** est un type de **thérapie**

Objectifs et méthodologie

Objectifs

Intégration de la morphologie dans le processus d'acquisition automatique de ressources lexicales à partir de textes

Il est nécessaire de disposer de ressources morphologiques

Méthodologie et matériel

- Travail sur corpus
- Langue de spécialité
- Apprentissage et approche statistique
- Indépendance aux langues

Schéma global

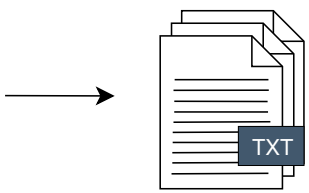
Données textuelles

Internet

Schéma global

Données textuelles

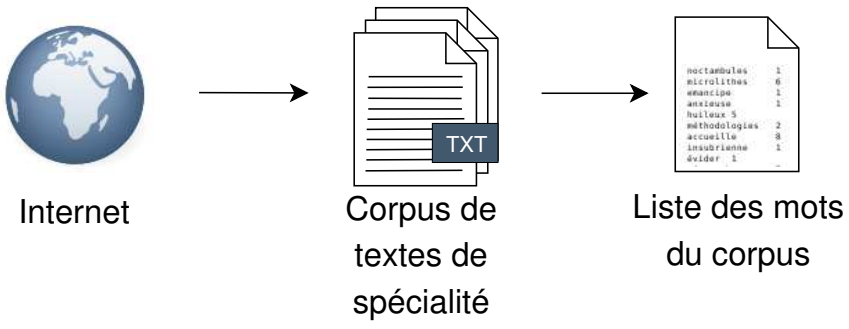
Internet



Corpus de textes de spécialité

Schéma global

Données textuelles



Internet

Corpus de textes de spécialité

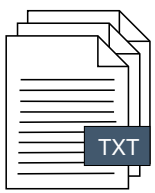
Liste des mots du corpus

Schéma global

Données textuelles

Analyse morphologique non supervisée

Internet



Corpus de textes de spécialité

Liste des mots du corpus

Schéma global

Données textuelles

Analyse morphologique non supervisée

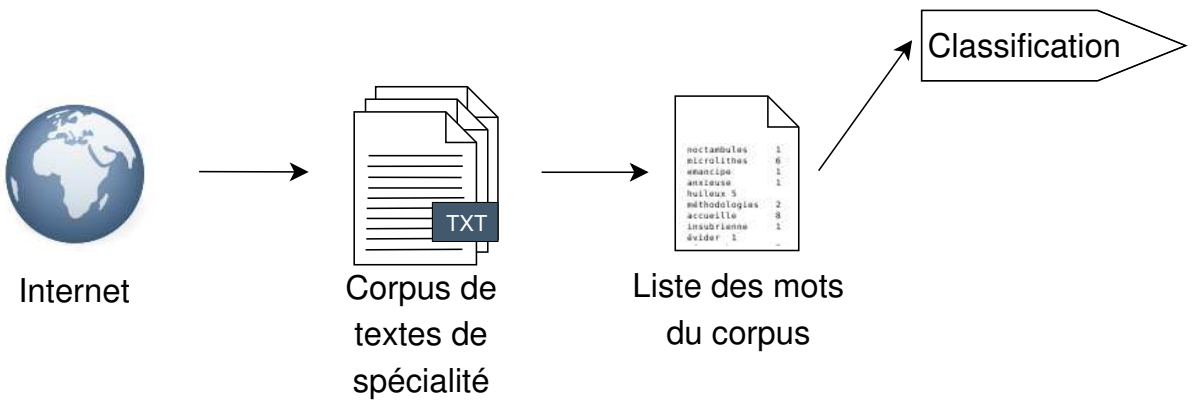


Schéma global

Données textuelles

Analyse morphologique non supervisée

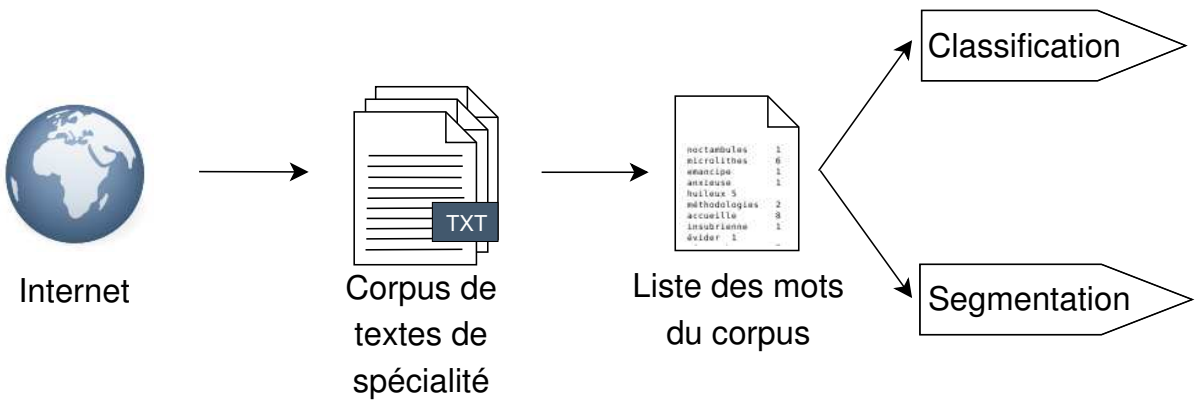


Schéma global

Données textuelles	Analyse morphologique non supervisée	Applications

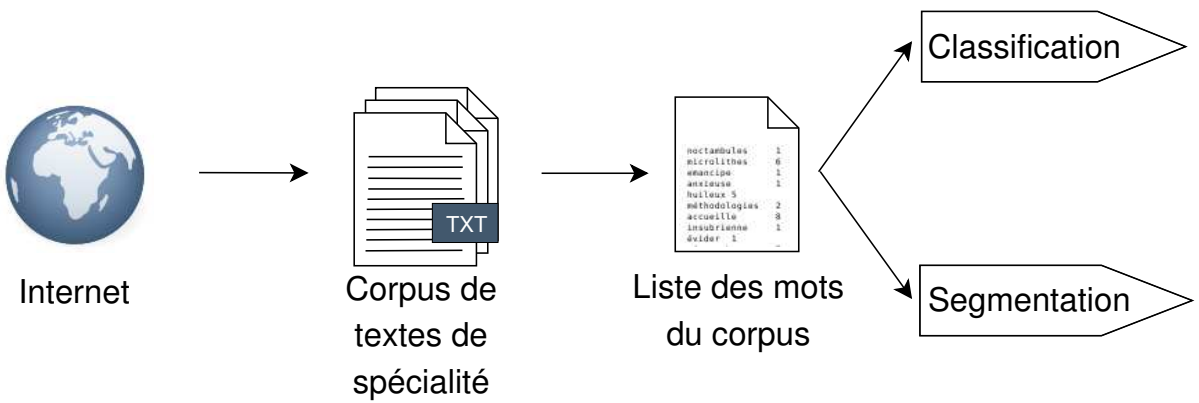


Schéma global

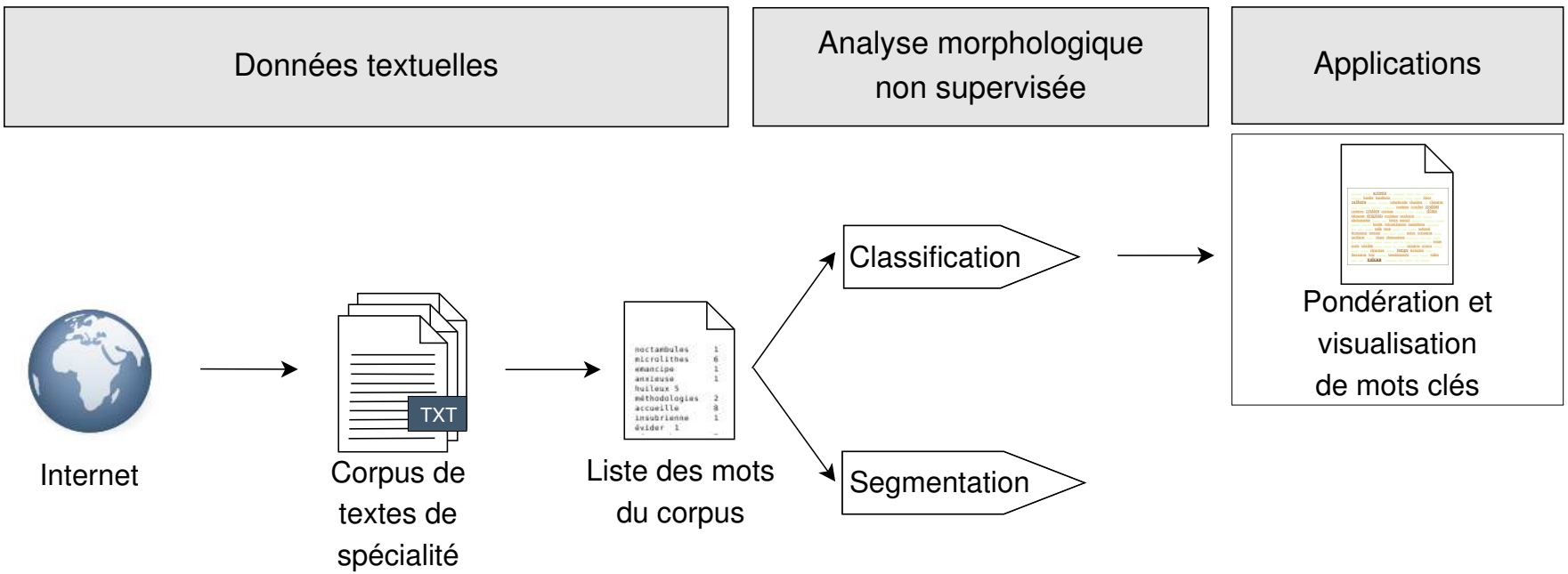
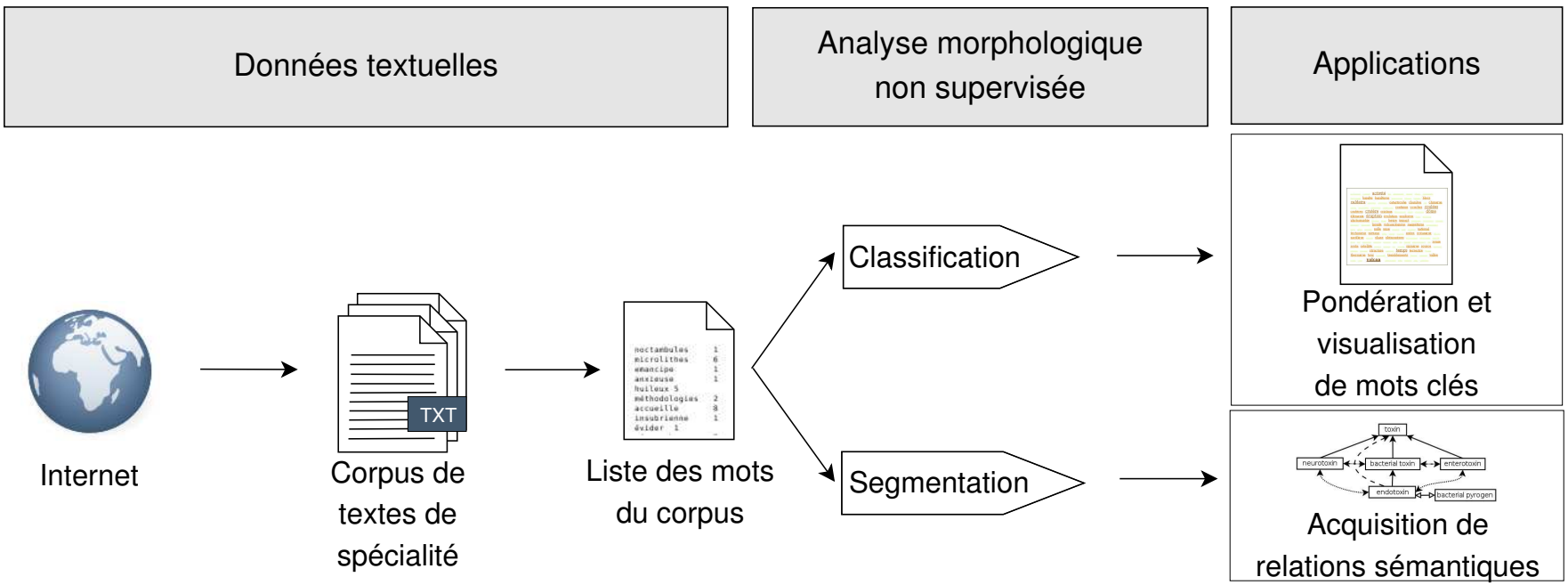


Schéma global



Plan

Plan

Travail sur corpus

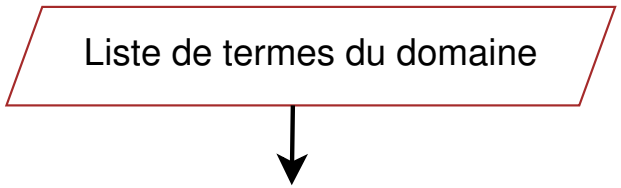
Corpus existant

Surtout généralistes, inadaptés au vocabulaire technique

Construction de corpus

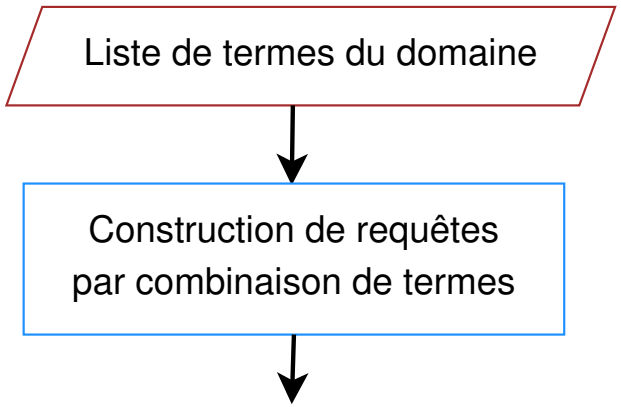
- Manuelle : processus long
- Automatique :
  - Source : le Web
  - Outils : inspirés de l'approche BootCat [Baroni et Bernardini, 2004]

Collecte d'URLs



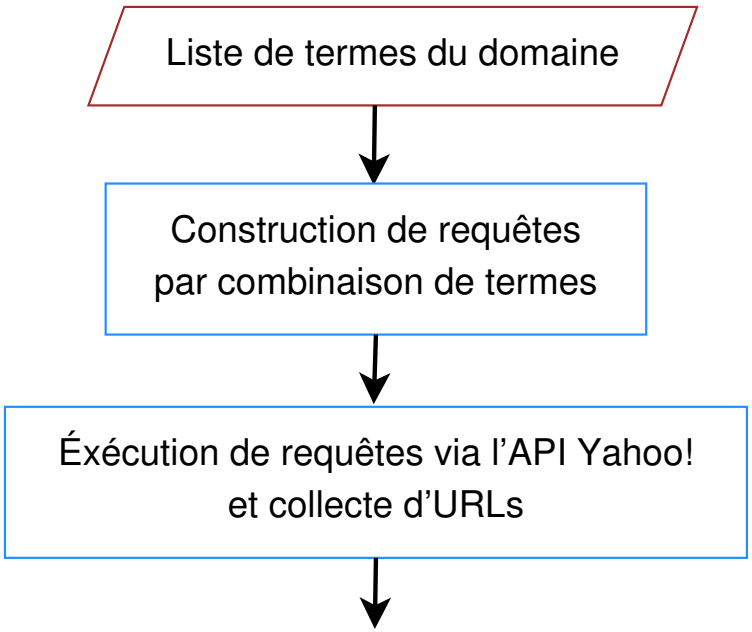
éruption
volcan
lave
magma
coulées
cratère
cendres
cône
caldeira
basalte
scories
dôme
ponces
pyroclastiques
...

Collecte d'URLs



éruption volcan lave
éruption volcan magma
magma coulées cratère
magma coulées cendres
magma coulées cône
coulées cône caldeira
cône caldeira basalte
cône caldeira scories
cône caldeira dôme
scories dôme ponces
scorie dôme pyroclastiques
...

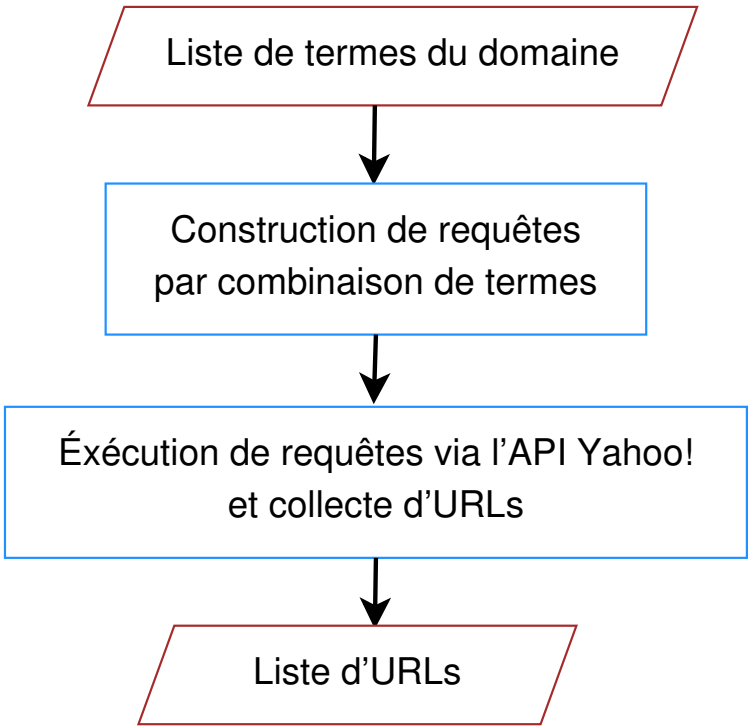
Collecte d'URLs



éruption volcan lave

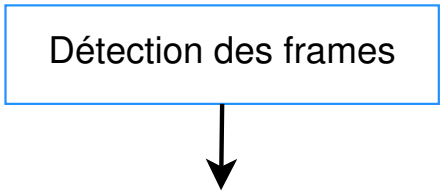
www.volcans.ch/pages/minute30_2002.html
www.runisland.com/volcan.html
fr.wikipedia.org/wiki/Volcan
www.volcanogeol.com/hawaii/lave.htm
users.skynet.be/lave.belgique
www.fournaise.info

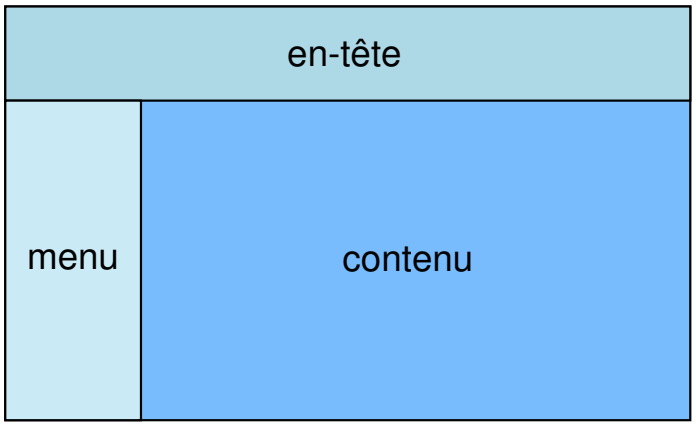
Collecte d'URLs



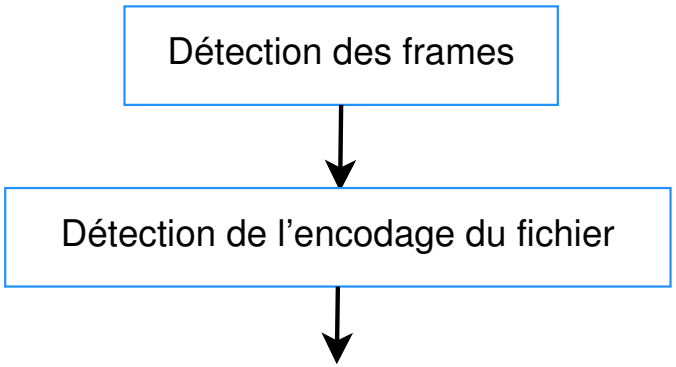
www.volcans.ch/pages/minute30_2002.html
www.runisland.com/volcan.html
fr.wikipedia.org/wiki/Volcan
www.volcanogeol.com/hawaii/lave.htm
users.skynet.be/lave.belgique
www.fournaise.info
www.ipgp.jussieu.fr/~aestp7/2002sicile.html
site.voila.fr/volcan
www.volcans.info/juillet_2001.htm
www.volcans2003.com/html/fiche/fiche1.htm
www.vulcania.com/fr/reperes-volcaniques-66.html
www.univ-ubs.fr/ecologie/volcanisme.html
www.volcan-actif.com/eruptions.htm
fr.rian.ru/russia/20050715/40914661.html
...

Collecte et pré-traitement des fichiers



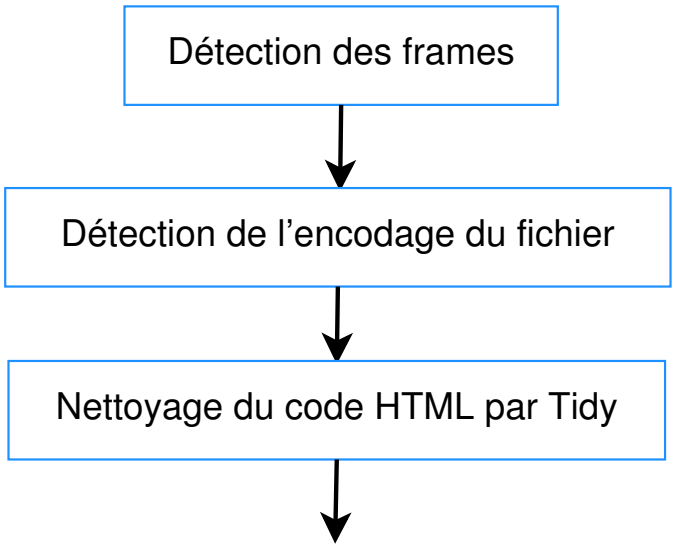


Collecte et pré-traitement des fichiers



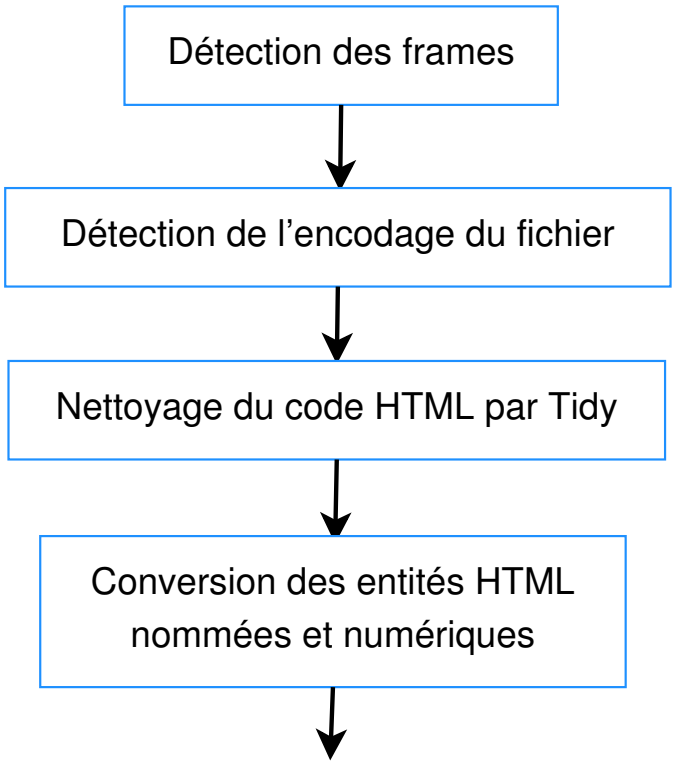

```
<html lang="fr">
<head>
<meta http-equiv="Content-Type"
content="text/html;
charset=iso-8859-1">
<title>...</title>
```

Collecte et pré-traitement des fichiers

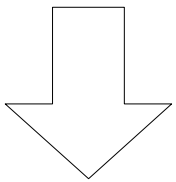



```
Warning: missing </h3> before <form>
Warning: inserting implicit <font>
Warning: discarding unexpected </font>
Warning: discarding unexpected </h3>
Warning: <spacer> is not approved by W3C
Warning: <spacer> is not approved by W3C
Warning: <spacer> is not approved by W3C
Warning: <spacer> is not approved by W3C
Warning: <spacer> is not approved by W3C
Error: <csobj> is not recognized!
221 warnings, 6 errors were found!
```

Collecte et pré-traitement des fichiers

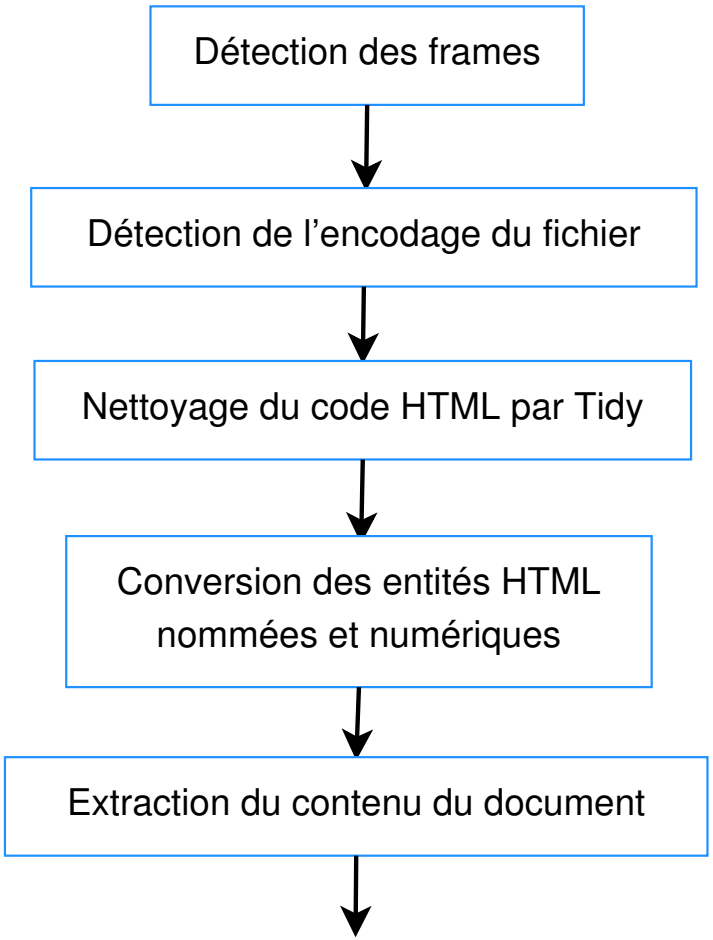



```
C'&eacute;tait, &#xE0; l'encoignure de la rue de la Michodière et de la rue Neuve-Saint-Augustin, un magasin de nouveaut&eacute;s dont les &eacute;talages &eacute;clataient en notes vives, dans la douce et p&#226;le journ&eacute;e d'octobre.
```



C'était, à l'encoignure de la rue de la Michodière et de la rue Neuve-Saint-Augustin, un magasin de nouveautés dont les étalages éclataient en notes vives, dans la douce et pâle journée d'octobre.

Collecte et pré-traitement des fichiers




encyclopaedic.net L'ENCYCLOPÉDIE FRANÇAISE

encyclopaedic.net || l'encyclopédie française > page d'accueil | index alphabétique

Une encyclopédie est un ouvrage où l'on tente de traiter de façon exhaustive de l'ensemble du savoir humain. Par extension, le mot désigne également un ouvrage qui traite systématiquement d'un domaine de connaissance en particulier. Le mot *encyclopédie* vient du grec ancien *enkuklios paideia* qui désignait l'éducation comprenant l'étude de toutes les sciences. Il devint *enkuklopaideia* par une erreur de transcription, puis au XVIe siècle les savants le latinisèrent en *encyclopaedia*.

Annonces Goooooogle — Votre publicité ici — Cet article chez Wikipedia

Droit algérien
Droit algérien des affaires Droit fiscal algérien
www.benlawyers.fr

La Suite OpenPortal - DIF
Outil intégré pour la gestion du Droit Individuel à la Formation
www.OpenPortal.fr

Ejecta

Par les **ejecta**, les scientifiques définnisent les objets rejeté par les volcans qui ne sont pas de la lave pendant certaines éruptions explosives qui fragmentent le magma et qui retombent plus ou moins verticalement hors du cratère.

Ces ejecta sont classés suivant leur taille dans un ordre croissant, cendres, lapilli, blocs et bombes sont des ejecta. La majorité de ceux-ci retombent alors en épousant une trajectoire parabolique. Par ailleurs, les particules les plus fines peuvent atteindre une altitude suffisante pour être entraînées par les courants aériens sur des distances considérables.

Chaque éruption volcanique est caractérisée par une catégorie particulière d'ejecta.Grace à cela, les vulcanologues peuvent définir les types de volcans. Ce sont ces produits qui, par leur accumulation, forment progressivement les cônes volcaniques.

voir aussi

Sismologie

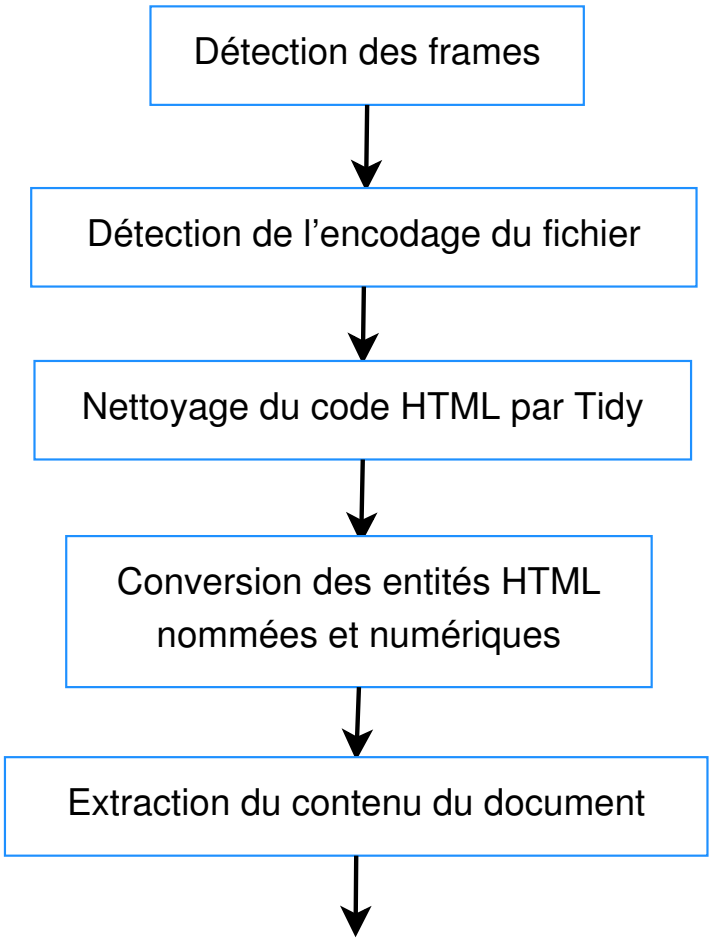
Annonces Google
Droit des affaires
Magazines sciences
Colloque droit
Droit bancaire
Droit

Source de l'article: Wikipedia. Le contenu est disponible selon les termes de la Licence de documentation libre GNU.

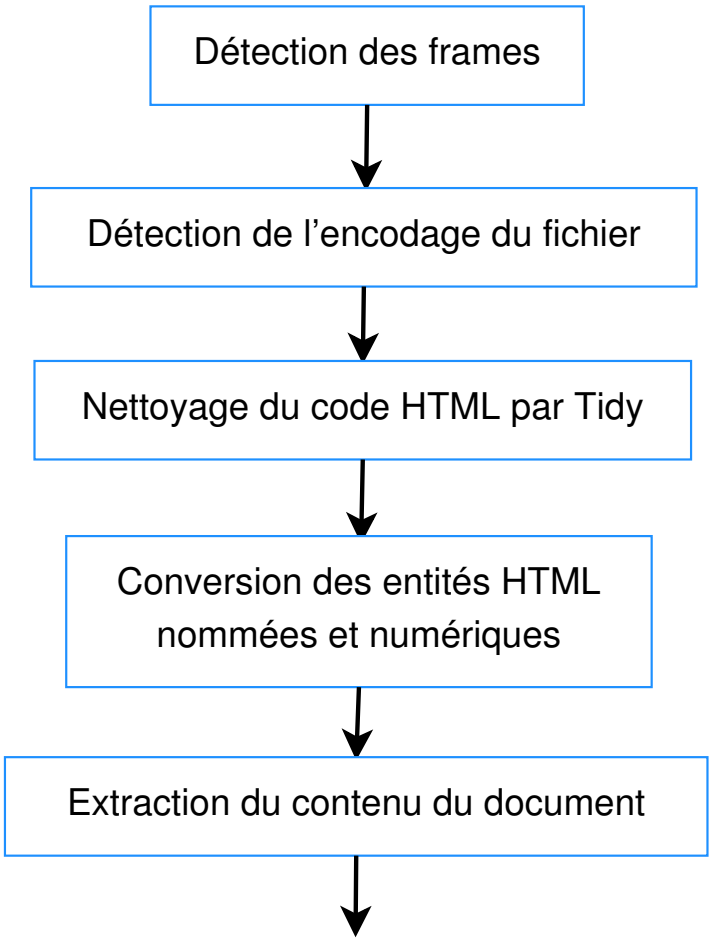
Sciences naturelles et mathématiques • Chimie • Écologie • Mathématiques • Physique • Sciences de la Terre • Sciences de l'Univers • Sciences de la Vie • Statistiques • Sciences humaines • Anthropologie • Archéologie • Éducation • Géographie • Histoire • Management • Langue et Linguistique • Musicologie • Pédagogie • Philosophie • Psychologie • Sciences cognitives • Sciences de l'information et des bibliothèques • Sociologie • Politique, droit et société • Associations et organismes • Commerce • Défense et sécurité • Droit • Économie • Entreprise • Famille • Gestion • Gestion de l'environnement • Métiers • Politique • Urbanisme • Religions et mythologies • Agnosticisme • Athéisme • Christianisme • Ésotérisme • Mysticisme • Mythologie • Religion • Sectes • Spiritualité • Théologie • Arts et culture • Art • Arts visuels • Arts du spectacle • Cinéma • Culture populaire • Danse • Littérature • Médias • Musique • Techniques et sciences appliquées • Aérospatiale • Agriculture • Architecture • Communication • Électricité • Électronique • Industrie • Informatique • Internet • Ingénierie • Médecine • Technologie • Télécommunications • Transport • Vie quotidienne et loisirs • Bricolage • Cuisine • Divertissement • Jardinage • Jeux • Nutrition • Santé • Sport • Tourisme • Divers • Liste des listes • Liste des pays du monde • Biographies • Arts appliqués

JAPANESE ENCYCLOPAEDIA | WORLD ENCYCLOPAEDIA | ENCYCLOPÉDIE FRANÇAISE | ENCICLOPEDIA ESPAÑOLA | ENCICLOPEDIA ITALIANA | ENCICLOPÉDIA PORTUGUESA | ENGLISH ENCYCLOPAEDIA

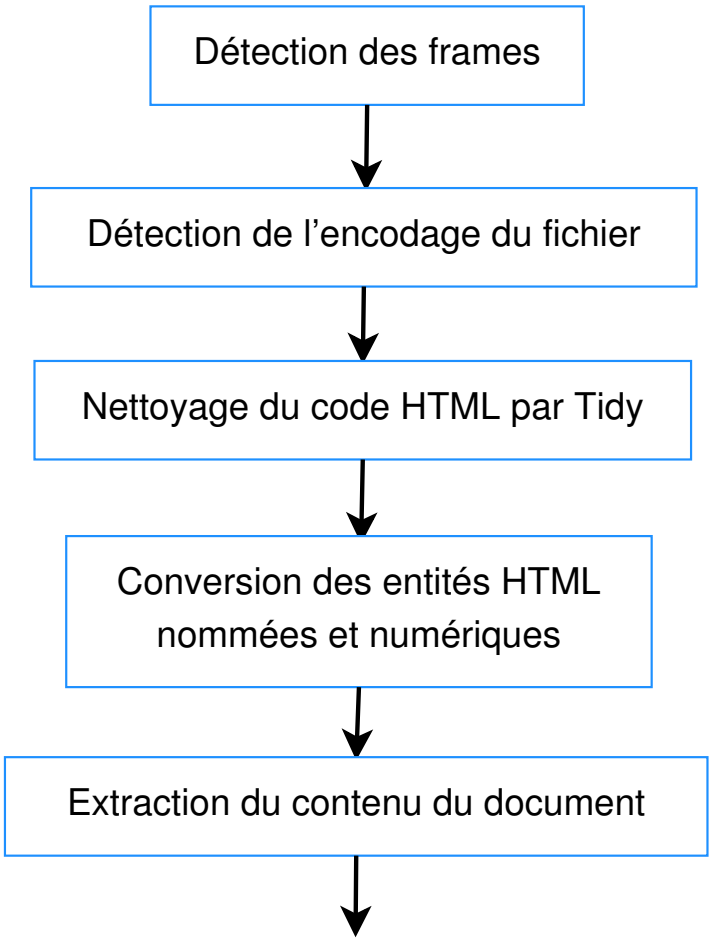
Collecte et pré-traitement des fichiers



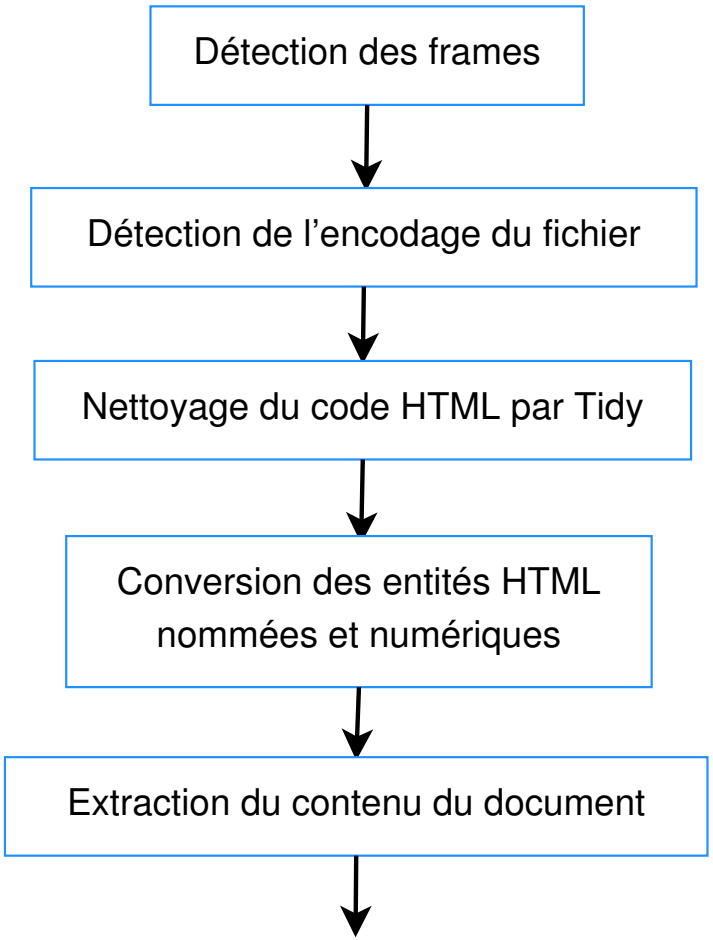
Collecte et pré-traitement des fichiers



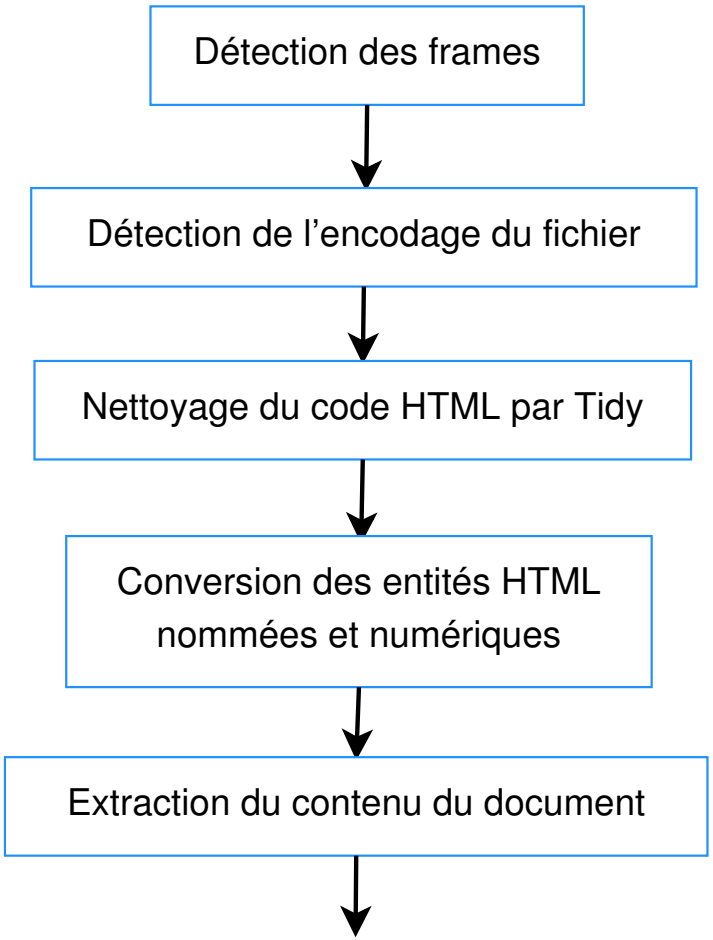
Collecte et pré-traitement des fichiers



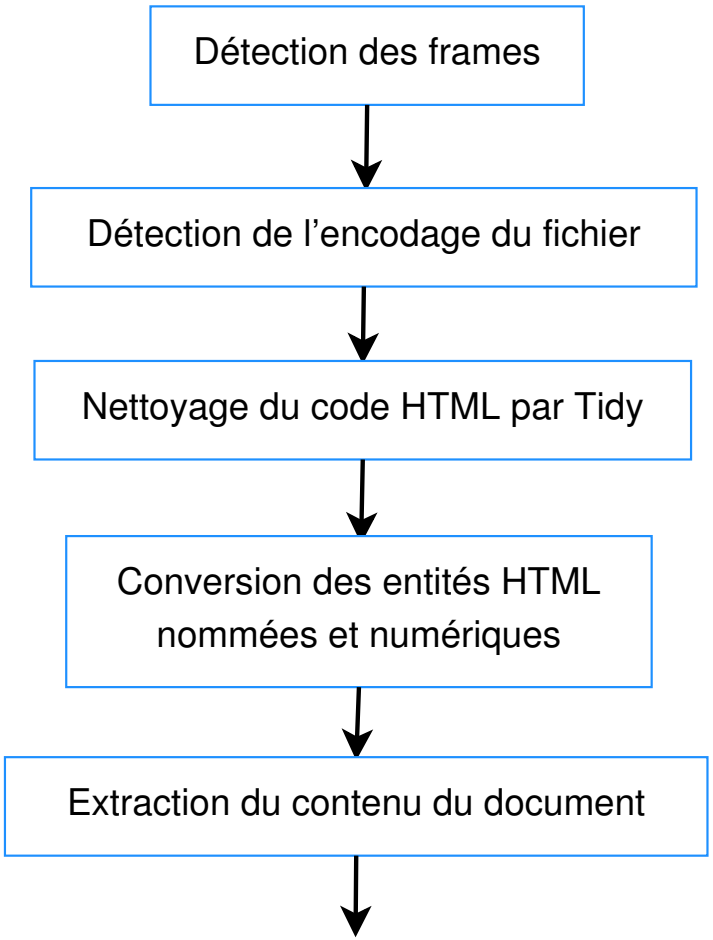
Collecte et pré-traitement des fichiers



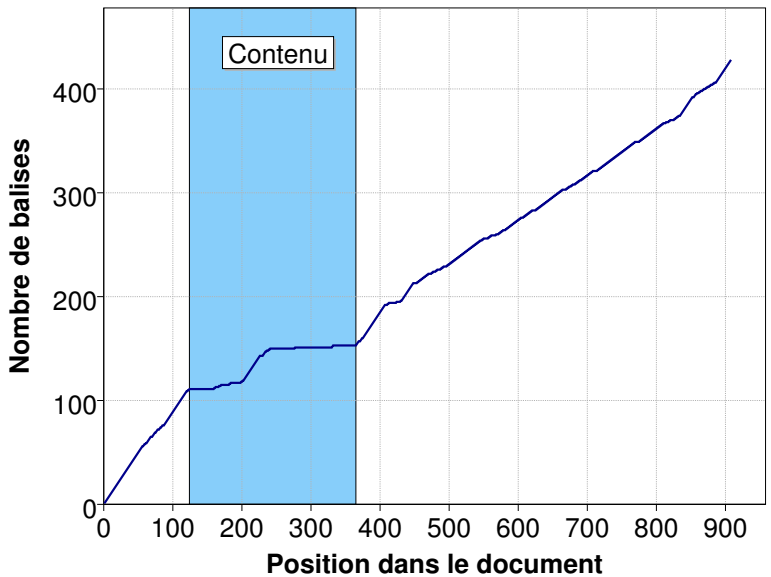
Collecte et pré-traitement des fichiers



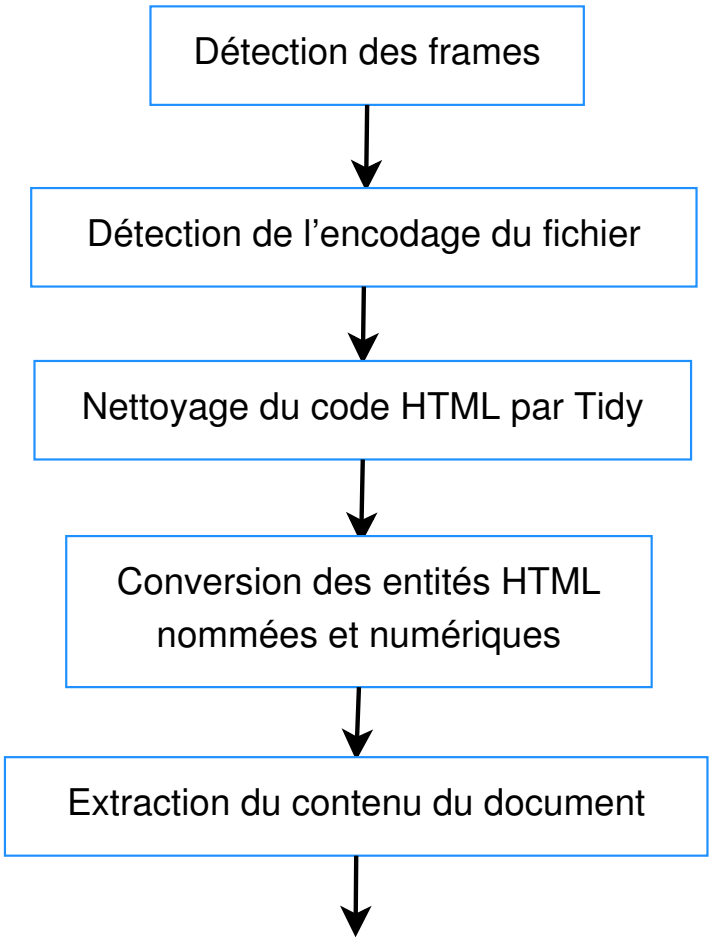
Collecte et pré-traitement des fichiers

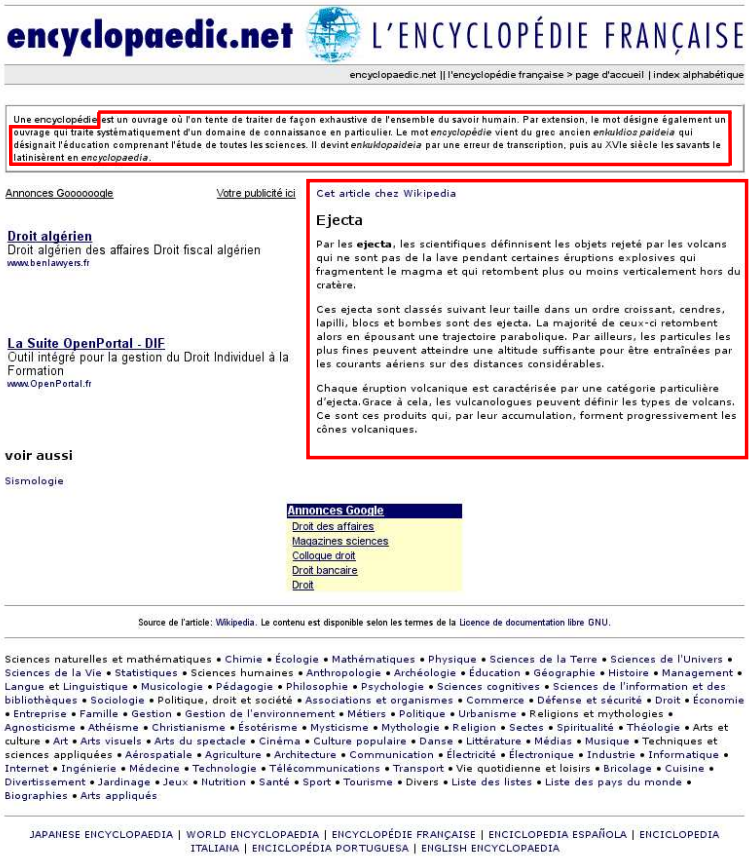


Méthode proposée par [Finn et al., 2001] : extraction de la sous-partie du document où la densité des mots est importante



Collecte et pré-traitement des fichiers




encyclopaedic.net L'ENCYCLOPÉDIE FRANÇAISE

encyclopaedic.net || l'encyclopédie française > page d'accueil | index alphabétique

Une encyclopédie est un ouvrage où l'on tente de traiter de façon exhaustive de l'ensemble du savoir humain. Par extension, le mot désigne également un ouvrage qui traite systématiquement d'un domaine de connaissance en particulier. Le mot *encyclopédie* vient du grec ancien *enkuklios paideia* qui désignait l'éducation comprenant l'étude de toutes les sciences. Il devint *enkuklopaideia* par une erreur de transcription, puis au XVIe siècle les savants le latinisèrent en *encyclopaedia*.

Annonces Goooooogle — Votre publicité ici

Droit algérien
Droit algérien des affaires Droit fiscal algérien
www.benlawyers.fr

La Suite OpenPortal - DIF
Outil intégré pour la gestion du Droit Individuel à la Formation
www.OpenPortal.fr

Cet article chez Wikipedia

Ejecta

Par les **ejecta**, les scientifiques définnisent les objets rejeté par les volcans qui ne sont pas de la lave pendant certaines éruptions explosives qui fragmentent le magma et qui retombent plus ou moins verticalement hors du cratère.

Ces ejecta sont classés suivant leur taille dans un ordre croissant, cendres, lapilli, blocs et bombes sont des ejecta. La majorité de ceux-ci retombent alors en épousant une trajectoire parabolique. Par ailleurs, les particules les plus fines peuvent atteindre une altitude suffisante pour être entraînées par les courants aériens sur des distances considérables.

Chaque éruption volcanique est caractérisée par une catégorie particulière d'ejecta.Grace à cela, les vulcanologues peuvent définir les types de volcans. Ce sont ces produits qui, par leur accumulation, forment progressivement les cônes volcaniques.

voir aussi

Sismologie

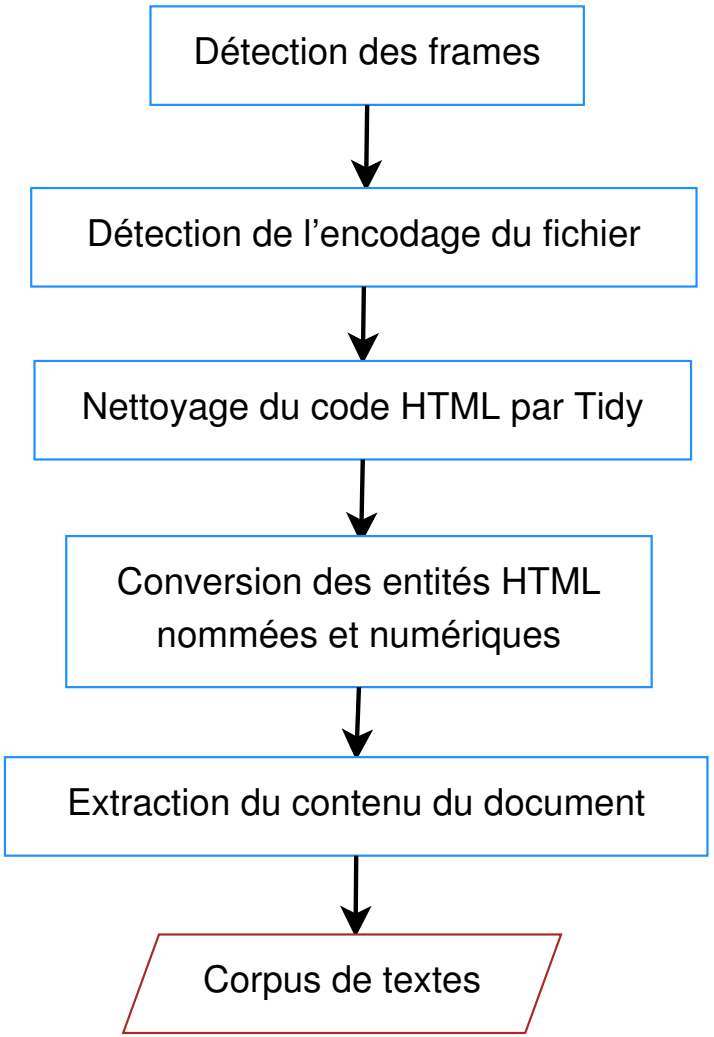
Annonces Google
Droit des affaires
Magazines sciences
Colloque droit
Droit bancaire
Droit

Source de l'article: Wikipedia. Le contenu est disponible selon les termes de la Licence de documentation libre GNU.

Sciences naturelles et mathématiques • Chimie • Écologie • Mathématiques • Physique • Sciences de la Terre • Sciences de l'Univers • Sciences de la Vie • Statistiques • Sciences humaines • Anthropologie • Archéologie • Éducation • Géographie • Histoire • Management • Langue et Linguistique • Musicologie • Pédagogie • Philosophie • Psychologie • Sciences cognitives • Sciences de l'information et des bibliothèques • Sociologie • Politique, droit et société • Associations et organismes • Commerce • Défense et sécurité • Droit • Économie • Entreprise • Famille • Gestion • Gestion de l'environnement • Métiers • Politique • Urbanisme • Religions et mythologies • Agnosticisme • Athéisme • Christianisme • Ésotérisme • Mysticisme • Mythologie • Religion • Sectes • Spiritualité • Théologie • Arts et culture • Art • Arts visuels • Arts du spectacle • Cinéma • Culture populaire • Danse • Littérature • Médias • Musique • Techniques et sciences appliquées • Aérospatiale • Agriculture • Architecture • Communication • Électricité • Électronique • Industrie • Informatique • Internet • Ingénierie • Médecine • Technologie • Télécommunications • Transport • Vie quotidienne et loisirs • Bricolage • Cuisine • Divertissement • Jardinage • Jeux • Nutrition • Santé • Sport • Tourisme • Divers • Liste des listes • Liste des pays du monde • Biographies • Arts appliqués

JAPANESE ENCYCLOPAEDIA | WORLD ENCYCLOPAEDIA | ENCYCLOPÉDIE FRANÇAISE | ENCICLOPEDIA ESPAÑOLA | ENCICLOPEDIA ITALIANA | ENCICLOPÉDIA PORTUGUESA | ENGLISH ENCYCLOPAEDIA

Collecte et pré-traitement des fichiers



Nombre de formes différentes

	Anglais	Français
Cancer du sein	86 149	46 898
Volcanologie	47 789	59 768

Analyse morphologique non supervisée

Deux approches

1. Segmentation : découpage des mots en segments morphémiques étiquetés
2. Classification : regroupement des mots dans des familles morphologiques

Contraintes

- Prise en compte des procédés de formation suivants :
  - flexion : carcinome carcinome**s**
  - dérivation : carcinome carcinom**ateux**
  - composition : carcinome **hépat**o**carcinom**e
- Méthode utilisable pour d'autres langues que l'anglais et le français et pour divers domaines

Analyse morphologique par segmentation

Données

Liste de mots

Étapes

1. Apprentissage de préfixes et de suffixes
2. Acquisition de bases
3. Segmentation des mots par alignement et comparaison
4. Sélection de la meilleure segmentation

Apprentissage de préfixes et de suffixes [1]

Entrée

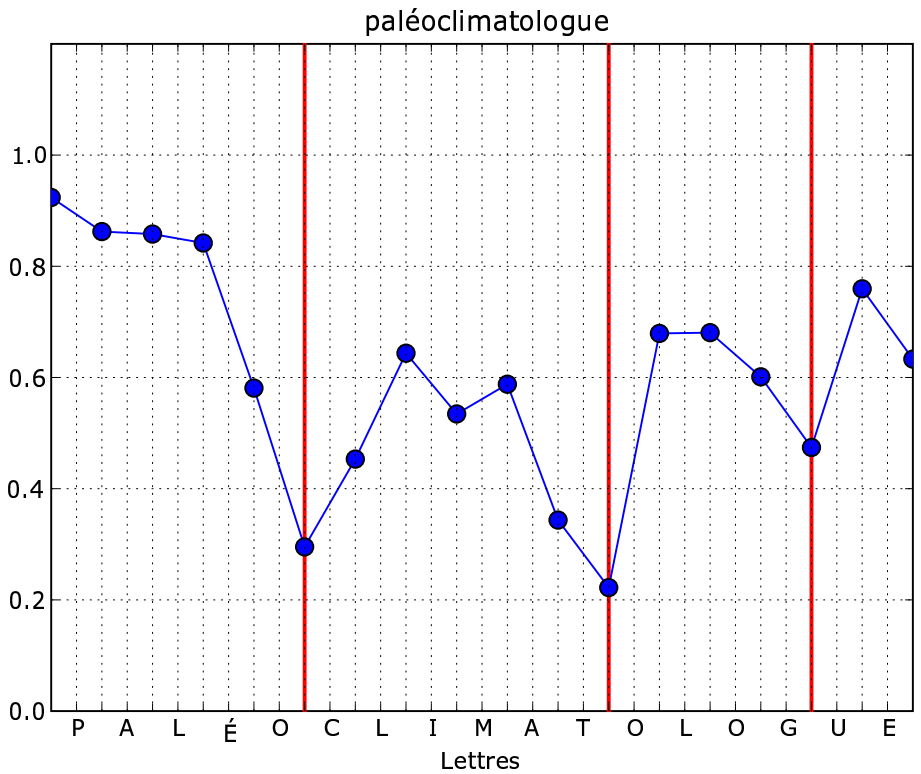
Mots les
plus longs

Apprentissage de préfixes et de suffixes [1]

Entrée

Mots les plus longs

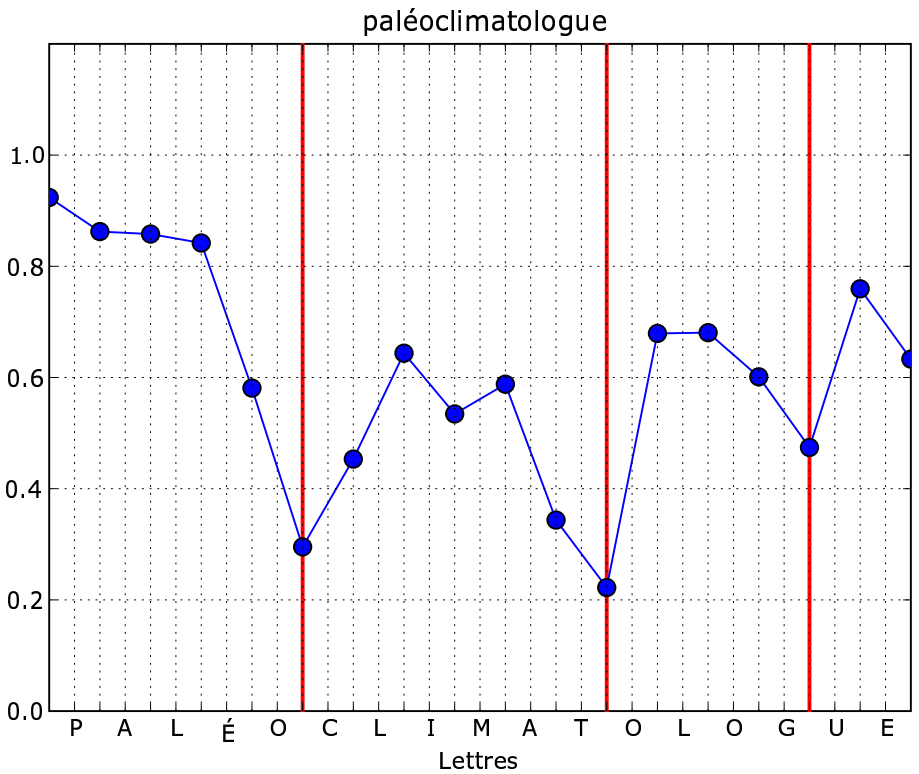
Localisation de segments



Apprentissage de préfixes et de suffixes [1]

Localisation de segments

Entrée
Mots les plus longs



Sortie
Segments

Apprentissage de préfixes et de suffixes [2]

Identification d'une base parmi les segments

	paléo	climat	olog	ue
fréquence	68	> 17 <	288	1 348
longueur	5	< 6 >	4	2

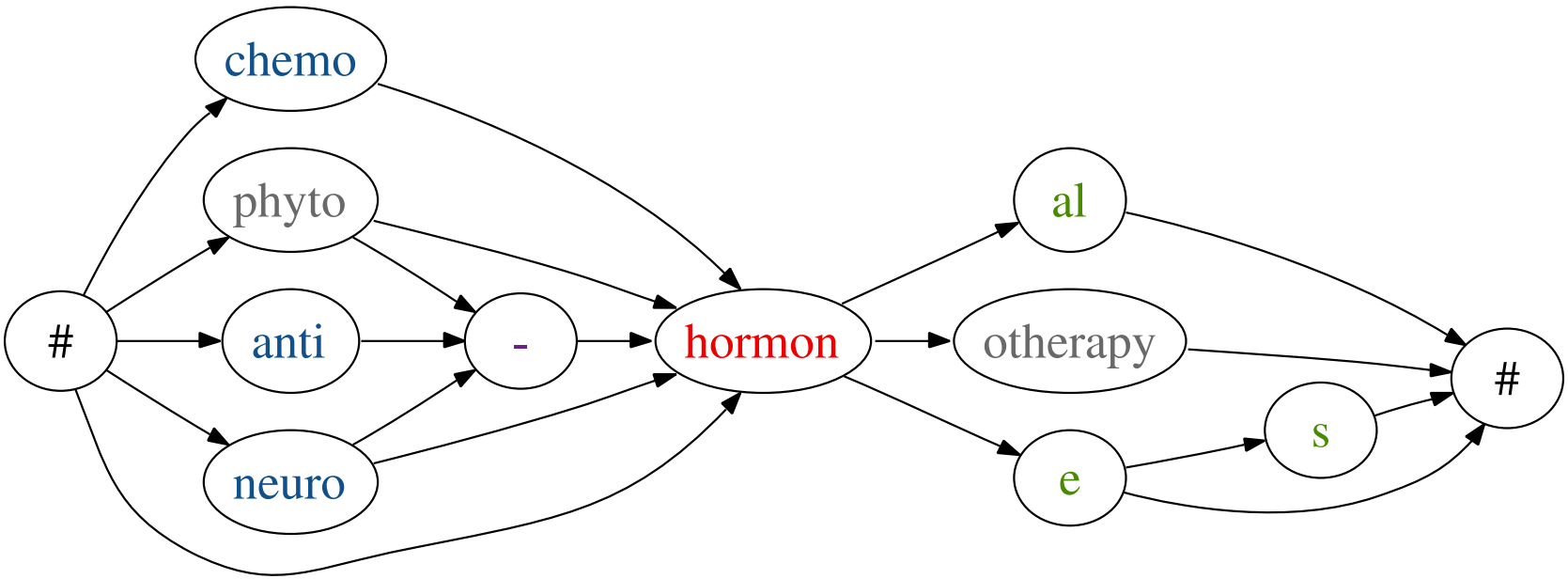
Préfixes et suffixes

paléo		s
		~~isation~~
		isés
	climat	s
paléo		ologue
ac		ation
~~dendro~~		ologie

Acquisition des bases

Retranchement des préfixes et des suffixes de tous les mots

Alignement des segments de mots [1]



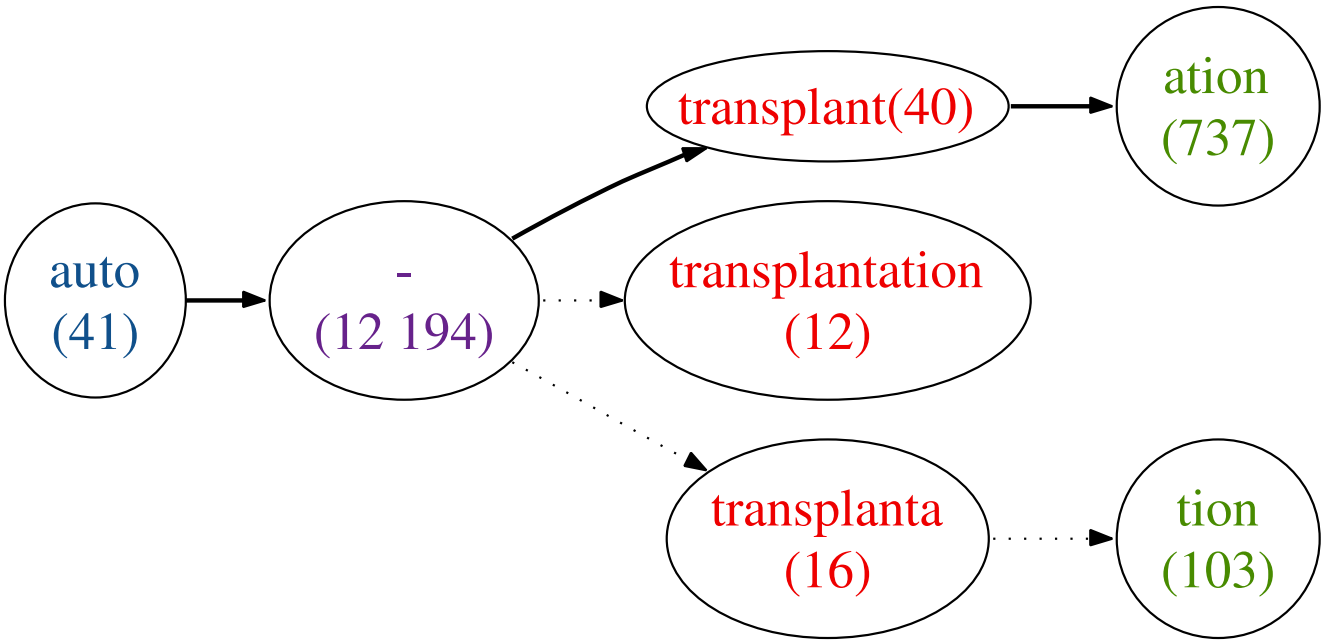
Alignement des segments de mots [2]

Validation des préfixes et suffixes inconnus

Mots	Suffixes connus A_1	Bases potentielles A_2	Nouveaux suffixes A_3
hormonal	-al		
hormonotherapy		-otherapy	
hormone	-e		
hormones			-es

$$\frac{|A_1| + |A_2|}{|A_1| + |A_2| + |A_3|} \geq a \text{ et } \frac{|A_1|}{|A_1| + |A_2|} \geq b$$

Sélection de la meilleure segmentation



Segmentation des mots absents du corpus d'apprentissage

- Sélection des segments qui minimisent le coût global
- Fonctions de coût utilisées :

$$coût_1(s_i) = -log\frac{f(s_i)}{\sum_i f(s_i)}$$

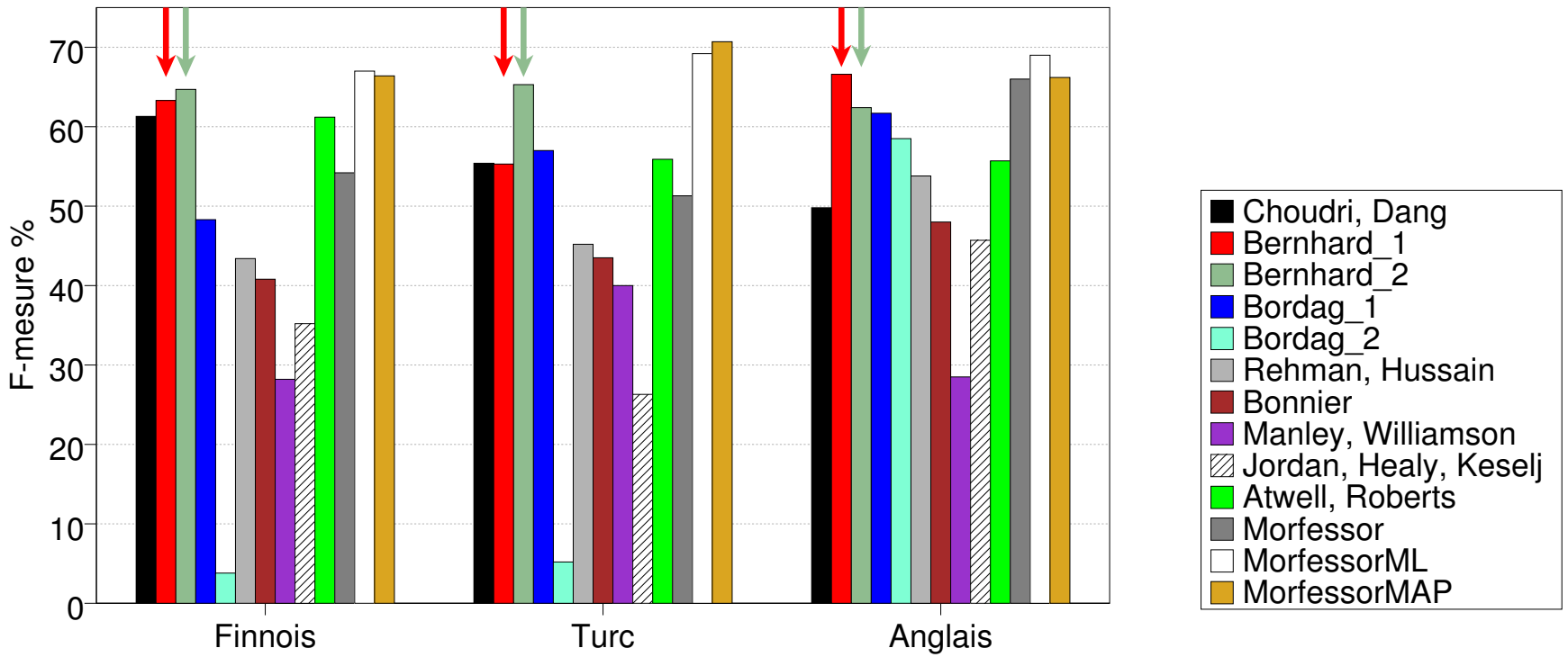
$$coût_2(s_i) = -log\frac{f(s_i)}{\max_i[f(s_i)]}$$

Exemples

Mots	**Segmentations**
allotransplantation	allo + transplantation
autotransplantation	auto + transplant + ation
post-transplant	post + - + transplant
retransplantation	re + transplant + ation
transplant	trans + pla + n + t
transplantation	trans + plant + ation
transplantations	trans + plant + ation + s
transplanté	trans + plant + é
transplantées	trans + plant + é + e + s
transplantés	trans + plant + é + s
transplants	trans + pla + n + t + s

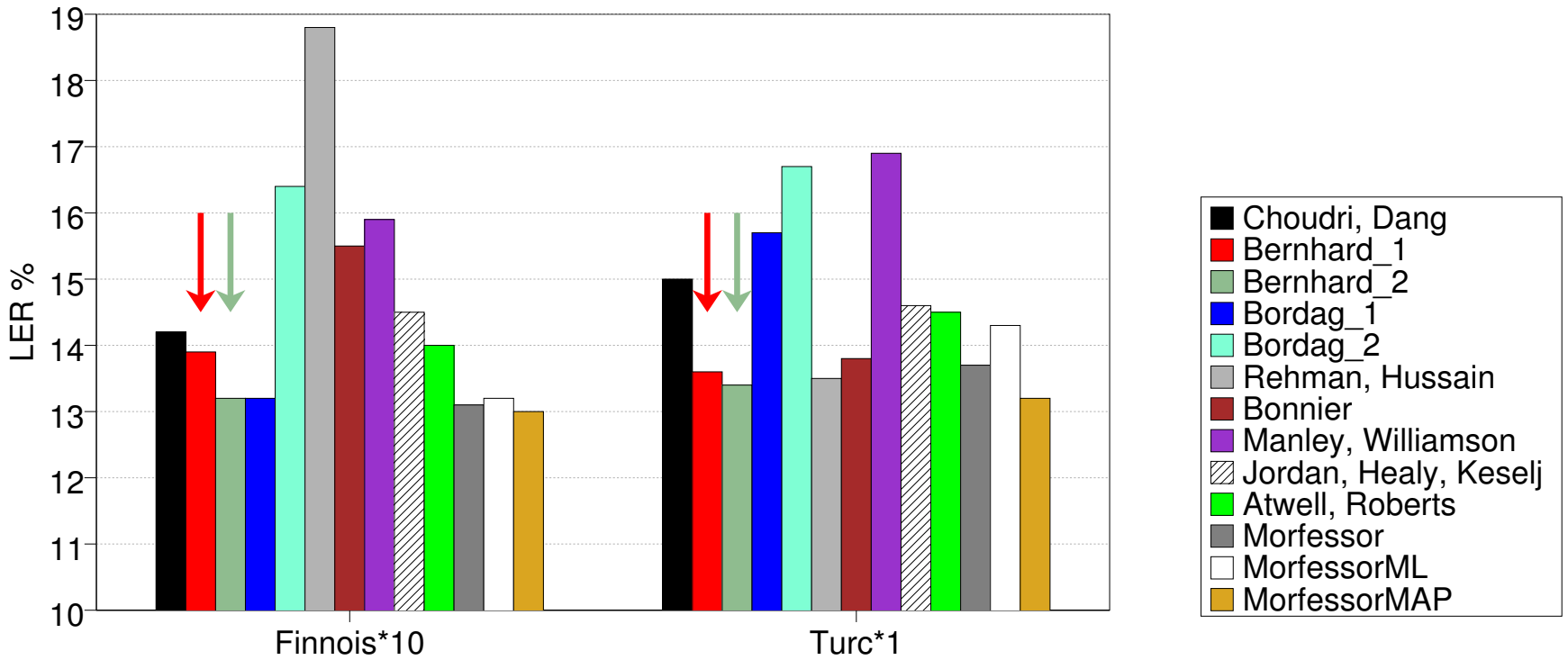
Évaluation 1 : Morpho Challenge

Compétition 1 : évaluation des segmentations



Évaluation 1 : Morpho Challenge

Compétition 2 : reconnaissance de la parole



Evaluation 2 : Synthèse de la parole

- Evaluation effectuée par V. Demberg
- Contexte : utilisation de la morphologie pour améliorer les résultats d'un système de conversion de graphèmes en phonèmes en allemand
- Résultats décevants : pas d'amélioration des résultats de la conversion
- Montre que les systèmes de segmentation morphologique non supervisés n'obtiennent pas encore une F-mesure suffisante

Évaluation 3 : Familles morphologiques

Familles de référence

- CELEX pour l'anglais
- Familles construites manuellement pour l'anglais et le français

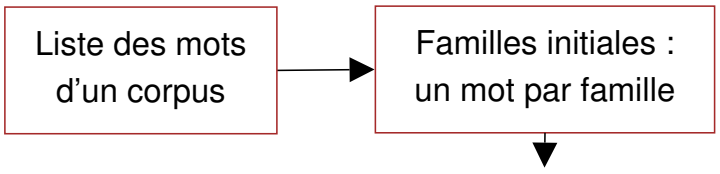
Mesure d'évaluation

Prise en compte du nombre d'éléments corrects, insérés et supprimés dans une famille morphologique par rapport aux familles de références.

Résultats obtenus

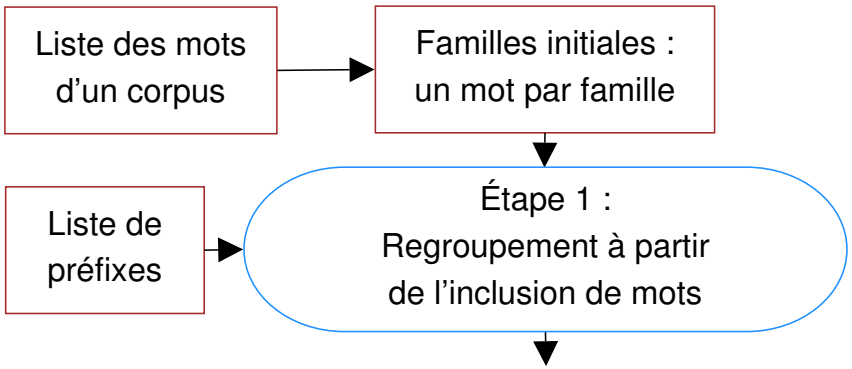
Proches des résultats de MorphoChallenge : F-mesure entre 60 et 70%.

Analyse morphologique par classification



subvolcaniques
sub-volcaniques
post-volcaniques
volcaniques
paléo-volcan
volcan
paléovolcanique
volcanique
subocéanique
océanique
sub-océaniques
océaniques
océan
océans

Analyse morphologique par classification



subvolcaniques
sub-volcaniques
post-volcaniques
volcaniques

paléo-volcan
volcan

paléovolcanique
volcanique

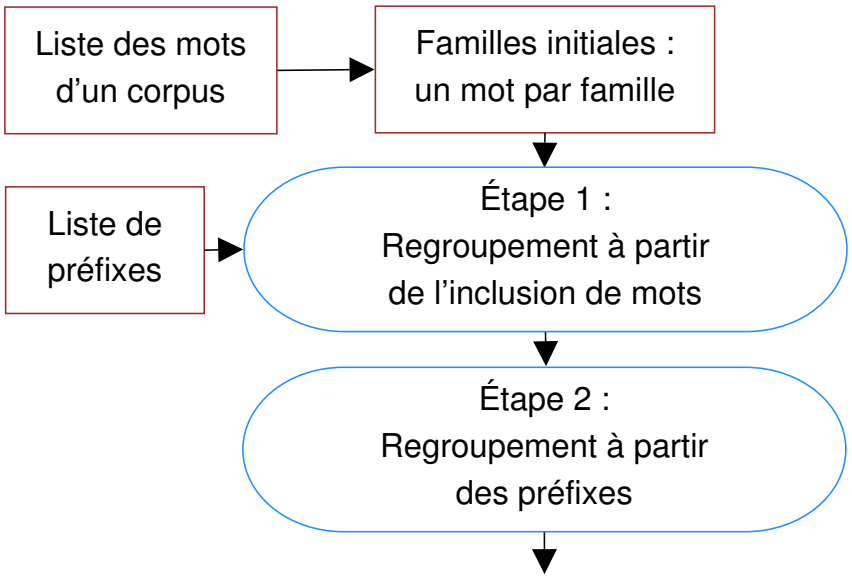
subocéanique
océanique

sub-océaniques
océaniques

océan

océans

Analyse morphologique par classification



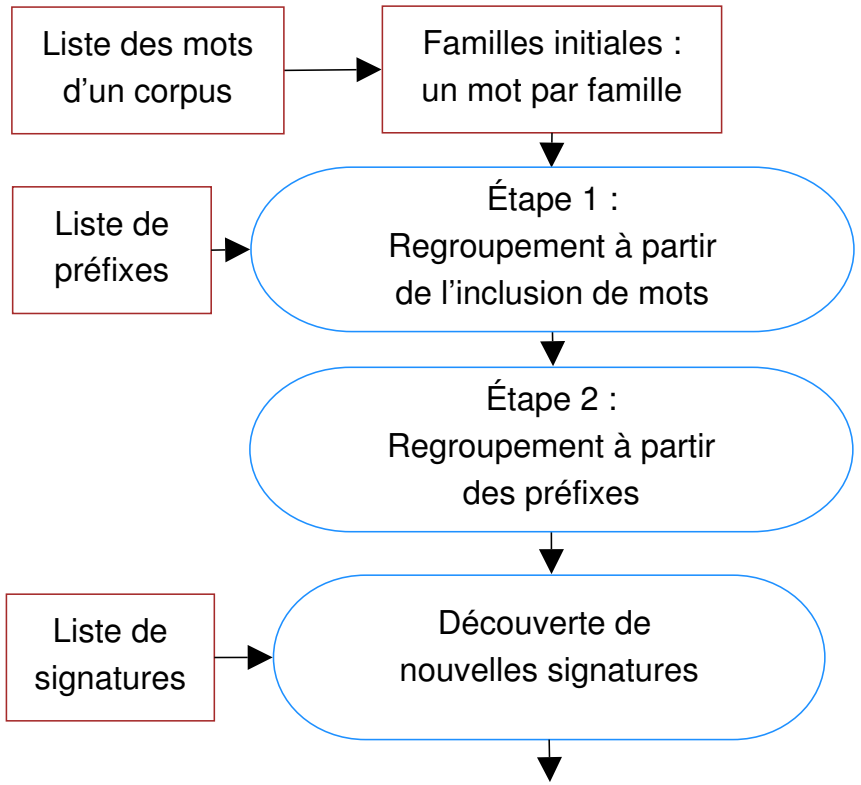
subvolcaniques
sub-volcaniques
post-volcaniques
volcaniques

paléo-volcan
volcan
paléovolcanique
volcanique

subocéanique
océanique
sub-océaniques
océaniques

océan
océans

Analyse morphologique par classification



subvolcaniques
sub-volcaniques
post-volcaniques
volcaniques

paléo-volcan
volcan
paléovolcanique
volcanique

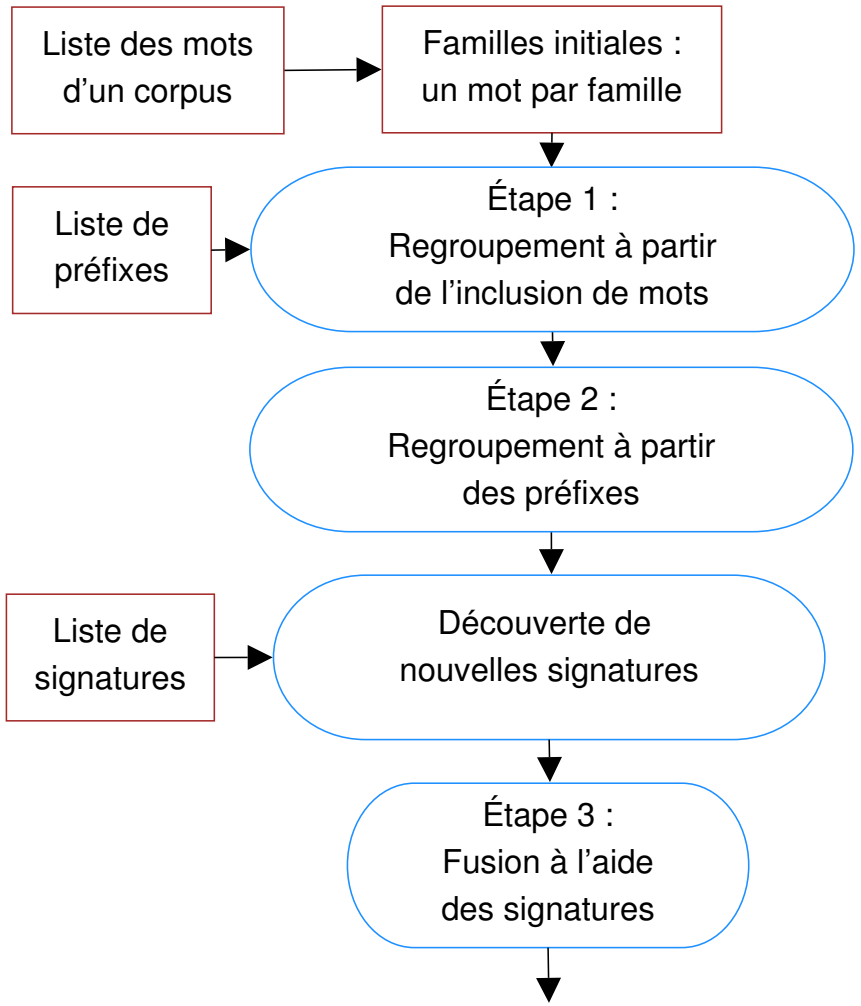
(ϵ,ique)

subocéanique
océanique
sub-océaniques
océaniques

(ϵ,s)

océan
océans

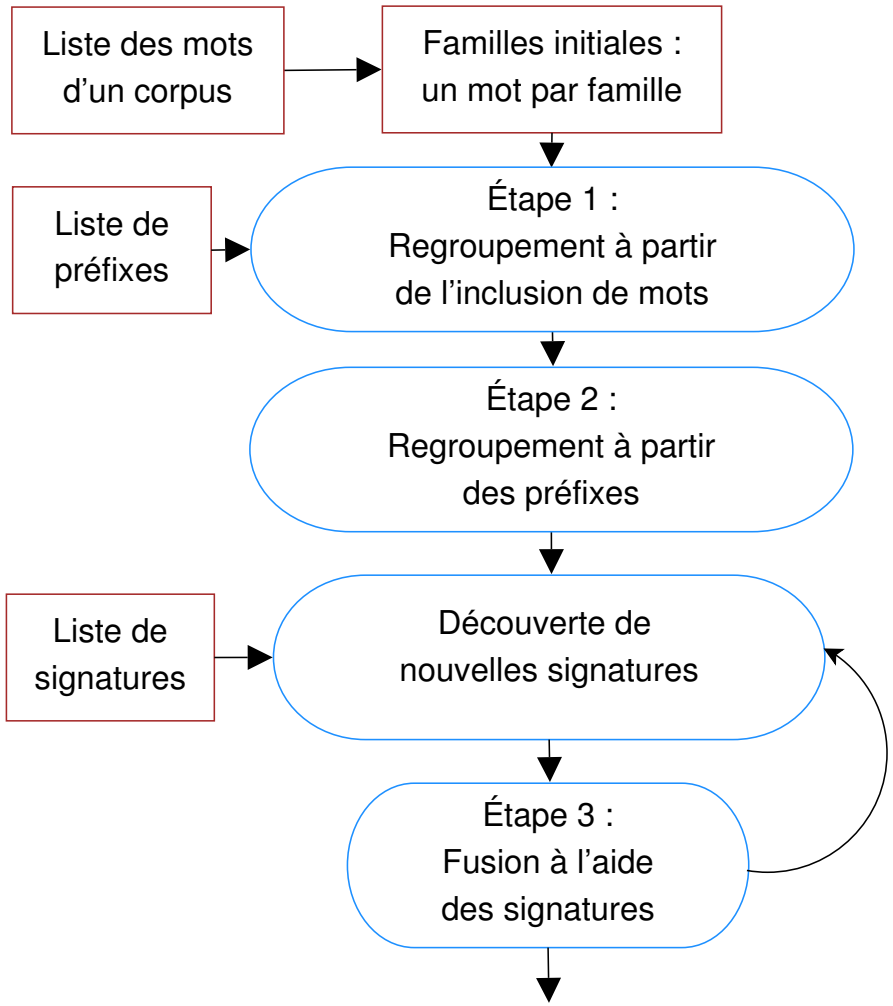
Analyse morphologique par classification



subvolcaniques
sub-volcaniques
post-volcaniques
volcaniques
paléo-volcan
volcan
paléovolcanique
volcanique

subocéanique
océanique
sub-océaniques
océaniques
océan
océans

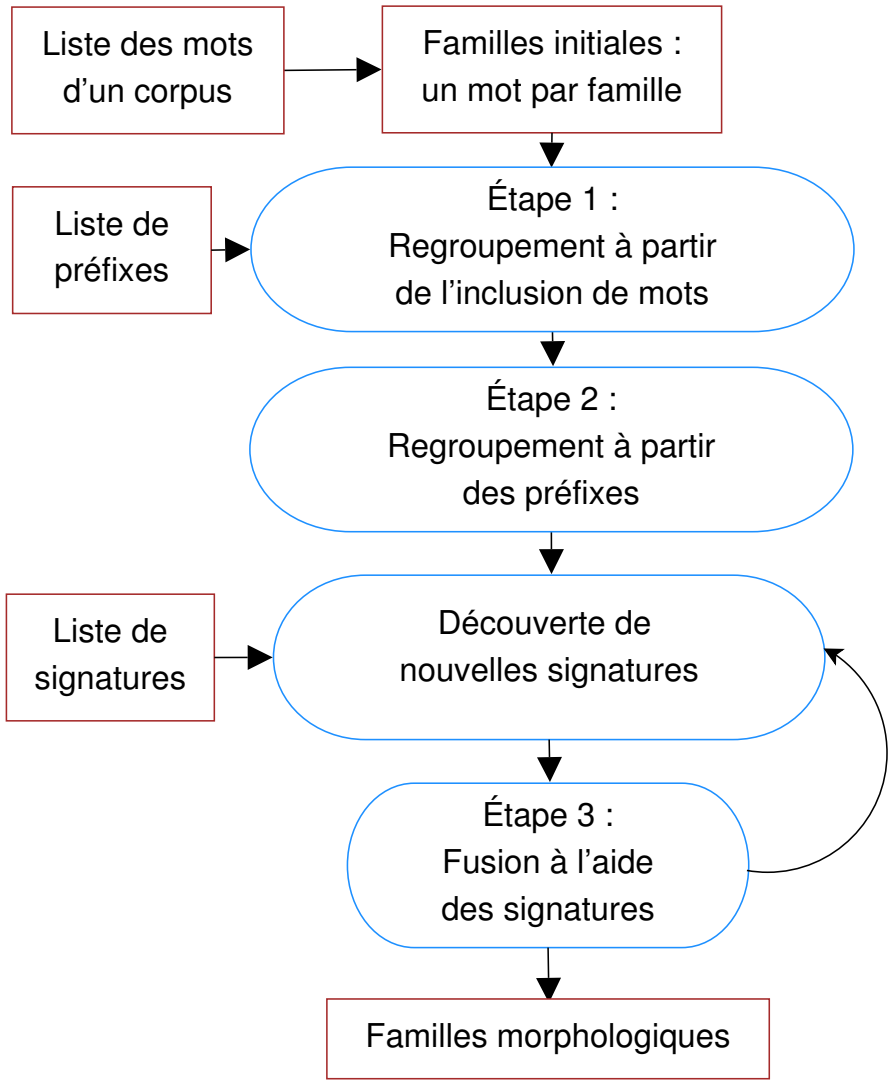
Analyse morphologique par classification



subvolcaniques	**(ε, ique)**
sub-volcaniques	**(ε, iques)**
post-volcaniques	**(ε, s)**
volcaniques	
paléo-volcan	
volcan	
paléovolcanique	
volcanique	

subocéanique	**(ε, ique)**
océanique	**(ε, iques)**
sub-océaniques	**(ε, s)**
océaniques	**(s, ique)**
océan	**(s, iques)**
océans	

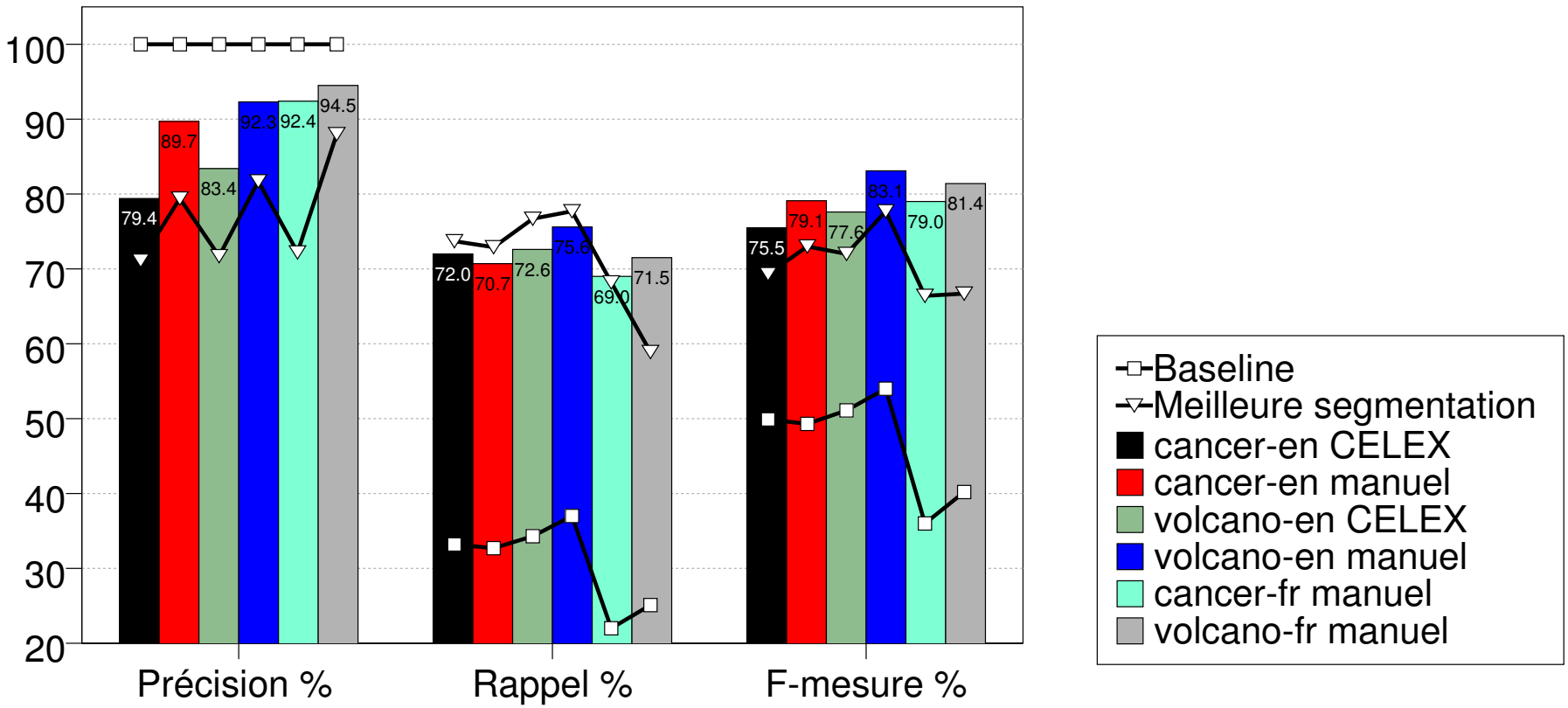
Analyse morphologique par classification



subvolcaniques
sub-volcaniques
post-volcaniques
volcaniques
paléo-volcan
volcan
paléovolcanique
volcanique

subocéanique
océanique
sub-océaniques
océaniques
océan
océans

Évaluation : Familles morphologiques



Analyse des résultats

Améliorations par rapport à l'analyse par segmentation

- Plus grande précision
- Doublement des consonnes en fin de radical
  [dimensio**nn**ement, dimensio**n**, dimensio**nn**ées, dimensio**n**s]
- Changements d'accentuation
  [crat**e**res, pseudocrat**è**re, intra-crat**è**re, pseudocrat**è**res, intra-crat**é**rique, crat**è**res, crat**e**rs, crat**e**r, crat**é**rique, pseudo-crat**è**res, crat**è**re, crat**e**re, intracrat**è**re, intracrat**é**rique]

Perspectives

- Évaluations complémentaires
- Classification multiple
- Déduction d'une segmentation des mots à partir de la classification

Plan

Pondération et visualisation de mots clés

Méthode

- **Mots clés** = mots qui décrivent le mieux le contenu d'un document ou d'un corpus
- Identification des familles de mots spécifiques au corpus étudié : combinaison d'indices structurels (familles morphologiques) et statistiques (fréquence)

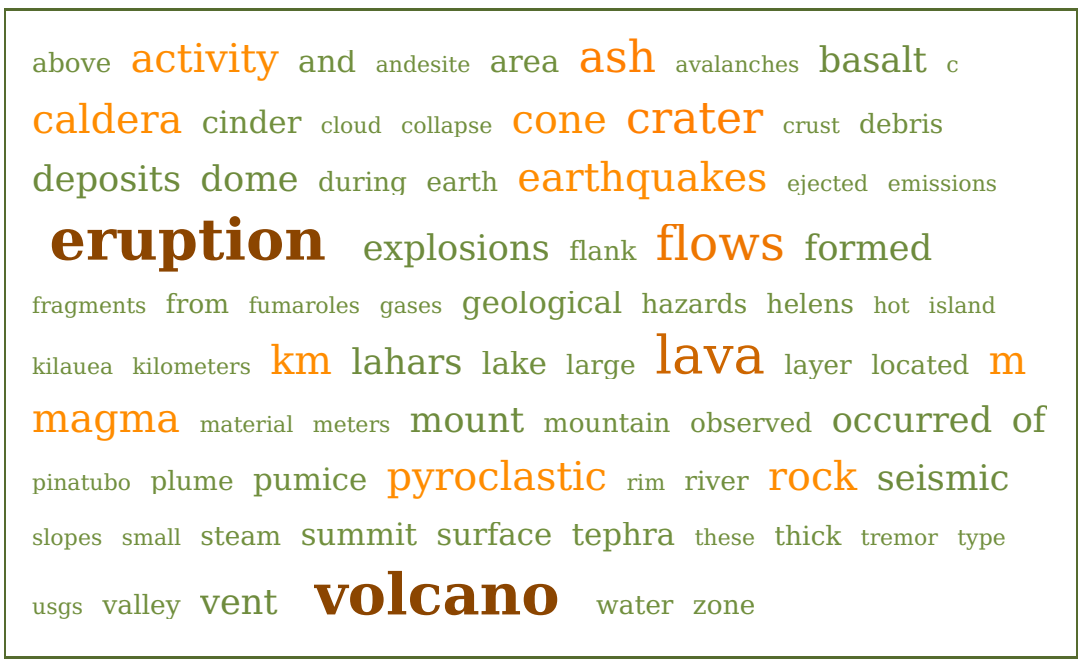
Mesures de pondération

- Fréquence de surface : nombre d'occurrences du mot dans le document ou le corpus considéré
- Fréquence cumulée : somme des fréquences de surface des mots appartenant à une même famille morphologique
- Comparaison des fréquences (de surface et cumulée) : log du rapport de vraisemblance

Visualisation des données

Liste pondérée au format HTML

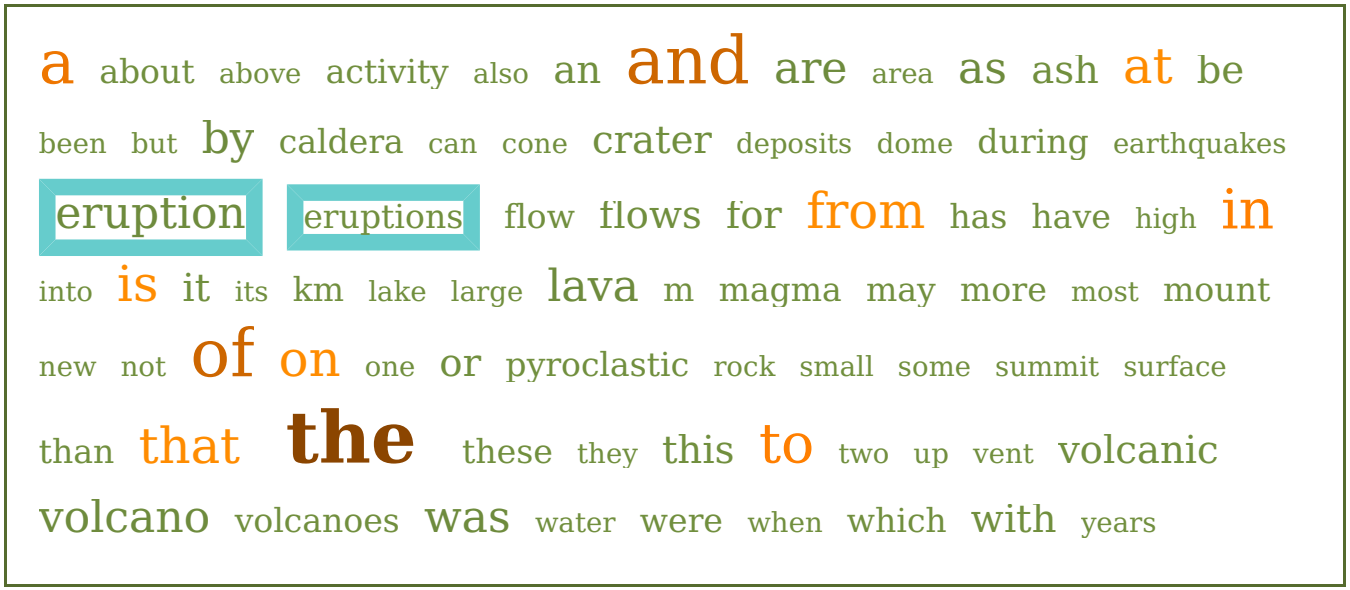
- **Liste pondérée** : la taille et la couleur d'un élément dépendent de son poids
- Carte des thématiques les plus importantes du corpus
- Représentation des familles par le mot le plus fréquent



andesictic	1
andesine	5
andesit	1
andesite	1336
andesite-based	1
andesite-dacite	15
andesites	132
andesitic	581
andesiting	1
basalt-andesite	4
basalt-andesite-dacite	3
dacitic-andesitic	4
	2084

Listes pondérées

Mots pondérés par la fréquence de surface



Listes pondérées

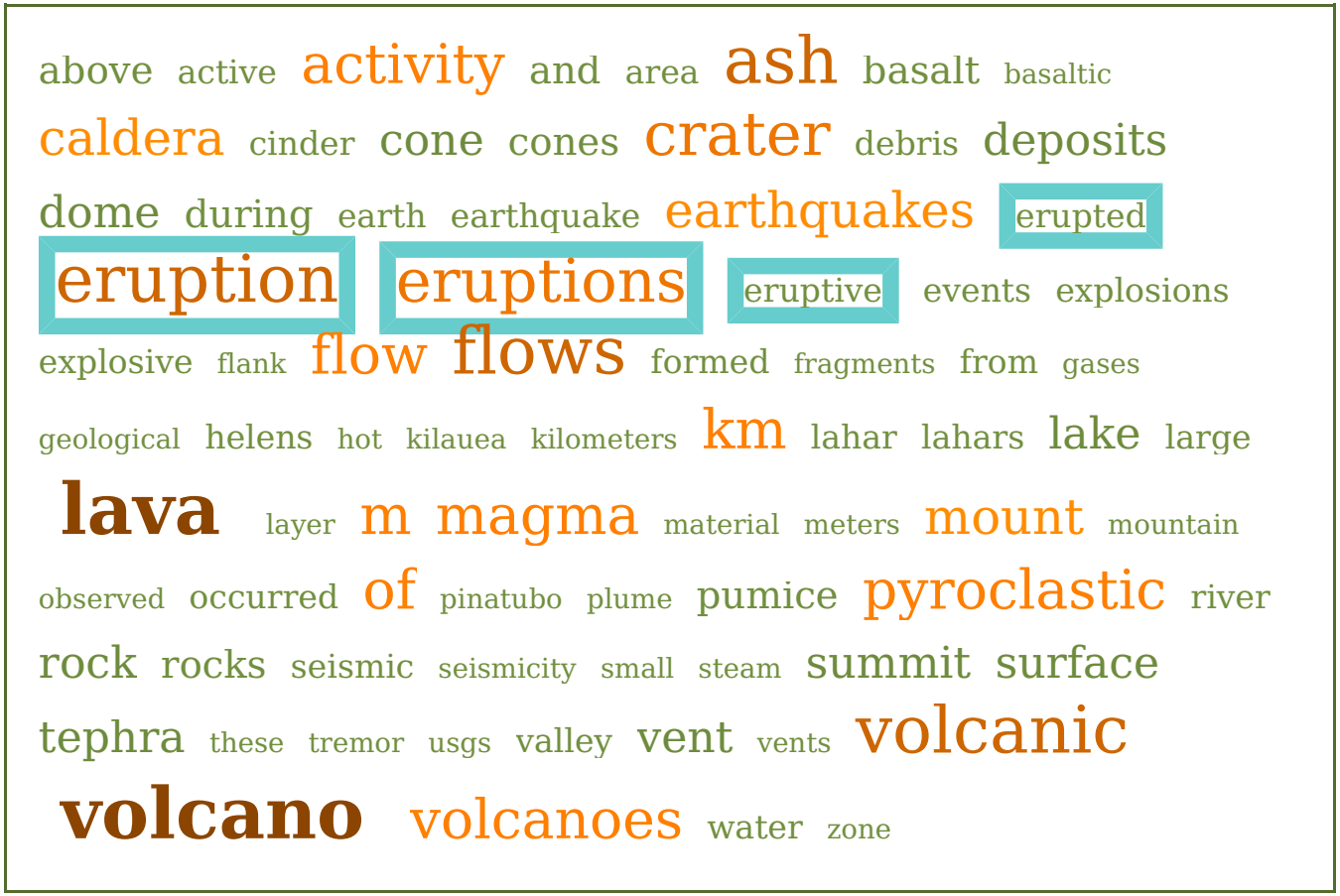
Familles pondérées par la fréquence cumulée

a about activity an and are area as ash at basalt be been but by caldera can cone continued crater deposits dome during earthquakes east eruption events explosions flows for formed from has have high in into is it km lahars lake large lava m magma may more most mount not occurred of on one or other produced pyroclastic report rock seismic small summit surface than that the these this time to vent volcano was were west which with years

erupt	926
erupted	2658
erupting	659
eruption	14923
eruption-induced	12
eruptions	9237
eruptive	2656
eruptives	6
eruptive-type	6
erupts	390
noneruption-induced	5
noneruptive	21
non-eruptive	23
noneruptively	3
posteruption	62
post-eruption	35
posteruptive	4
post-eruptive	5
preeruption	160
pre-eruption	89
preeruptive	18
pre-eruptive	20
	31947

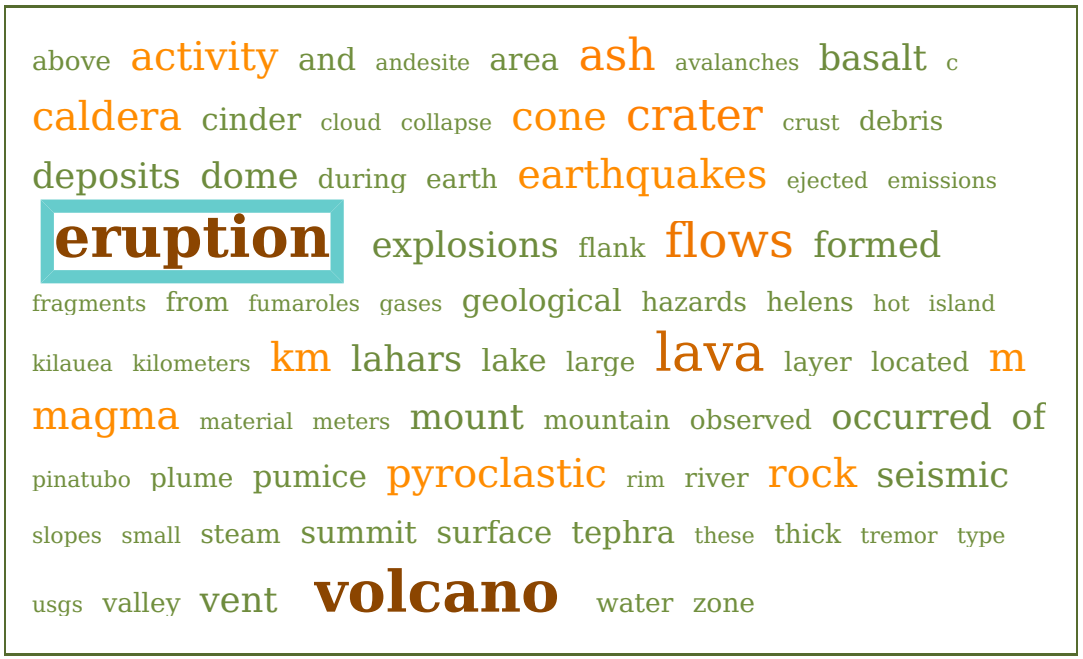
Listes pondérées

Mots pondérés par le log du rapport de vraisemblance



Listes pondérées

Familles pondérées par le log du rapport de vraisemblance



erupt	926
erupted	2658
erupting	659
eruption	14923
eruption-induced	12
eruptions	9237
eruptive	2656
eruptives	6
eruptive-type	6
erupts	390
noneruption-induced	5
noneruptive	21
non-eruptive	23
noneruptively	3
posteruption	62
post-eruption	35
posteruptive	4
post-eruptive	5
preeruption	160
pre-eruption	89
preeruptive	18
pre-eruptive	20
	31947

Acquisition de relations sémantiques

Quelles relations sémantiques ?

- Relations d'inclusion et d'identité :
  - Synonymie : livre - bouquin
  - Hyper-/Hyponymie (EST-UN) : chien - animal
  - Méronymie (PARTIE-DE) : bras - corps
- Antonymie : chaud - froid
- Co-hyponymie : chien - chat

Méthode

Relations structurelles basées sur les segments morphémiques

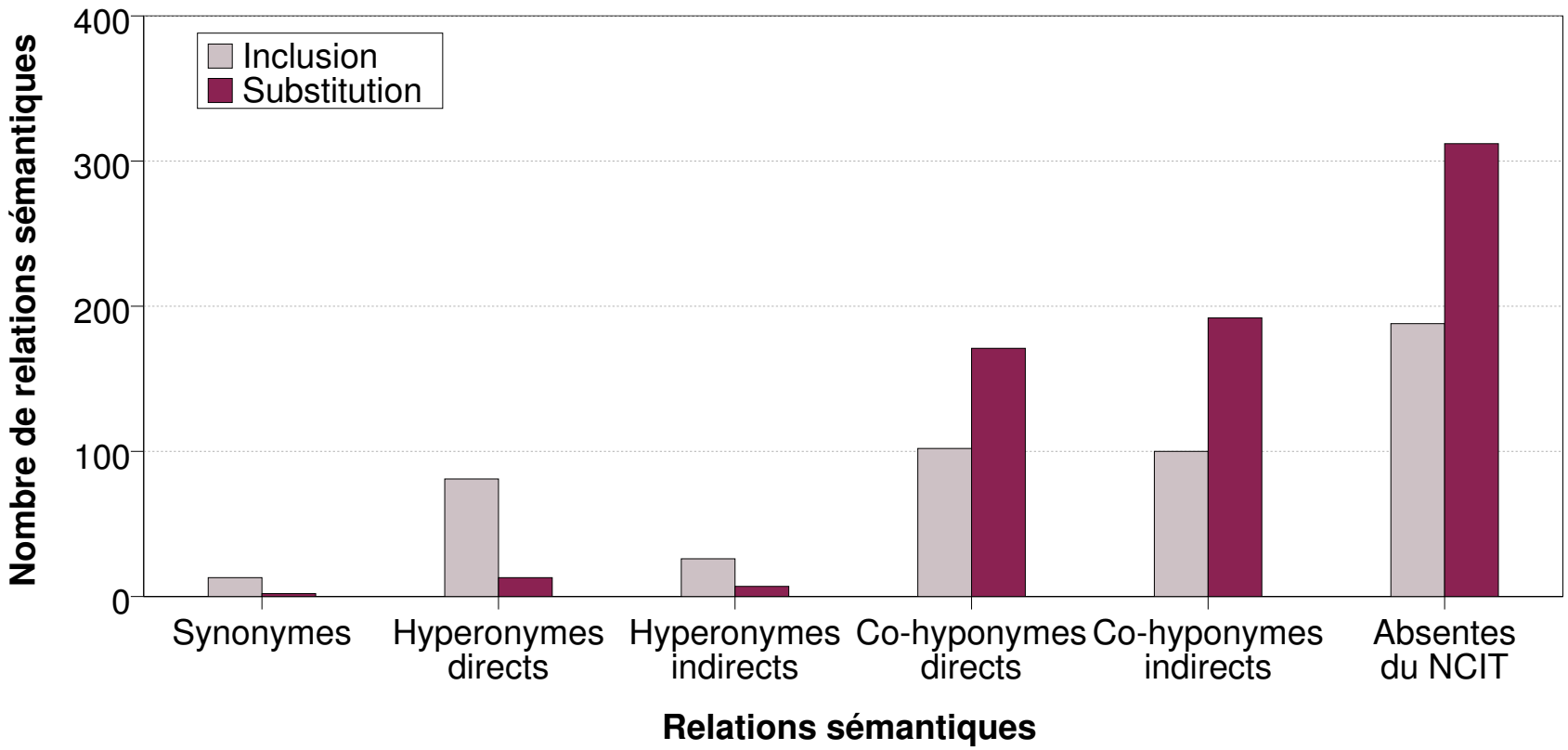
1. Inclusion
   a. Expansion gauche : *lymphedema* – *edema*
      [lymph + edema] – [edema]
   b. Insertion : *hepatosplenomegaly* – *hepatomegaly*
      [hepat + o + splen + o + mega + ly] – [hepat + o + mega + ly]
2. Substitution : *osteosarcoma* – *chondrosarcoma*
   [osteo + sarcoma] – [chondro + sarcoma]

Déduction de liens sémantiques

- Recherche de paires de mots liés par les relations structurelles précédentes
- Hypothèses :
  - Inclusion : hyper-/hyponymie
  - Substitution : co-hyponymie

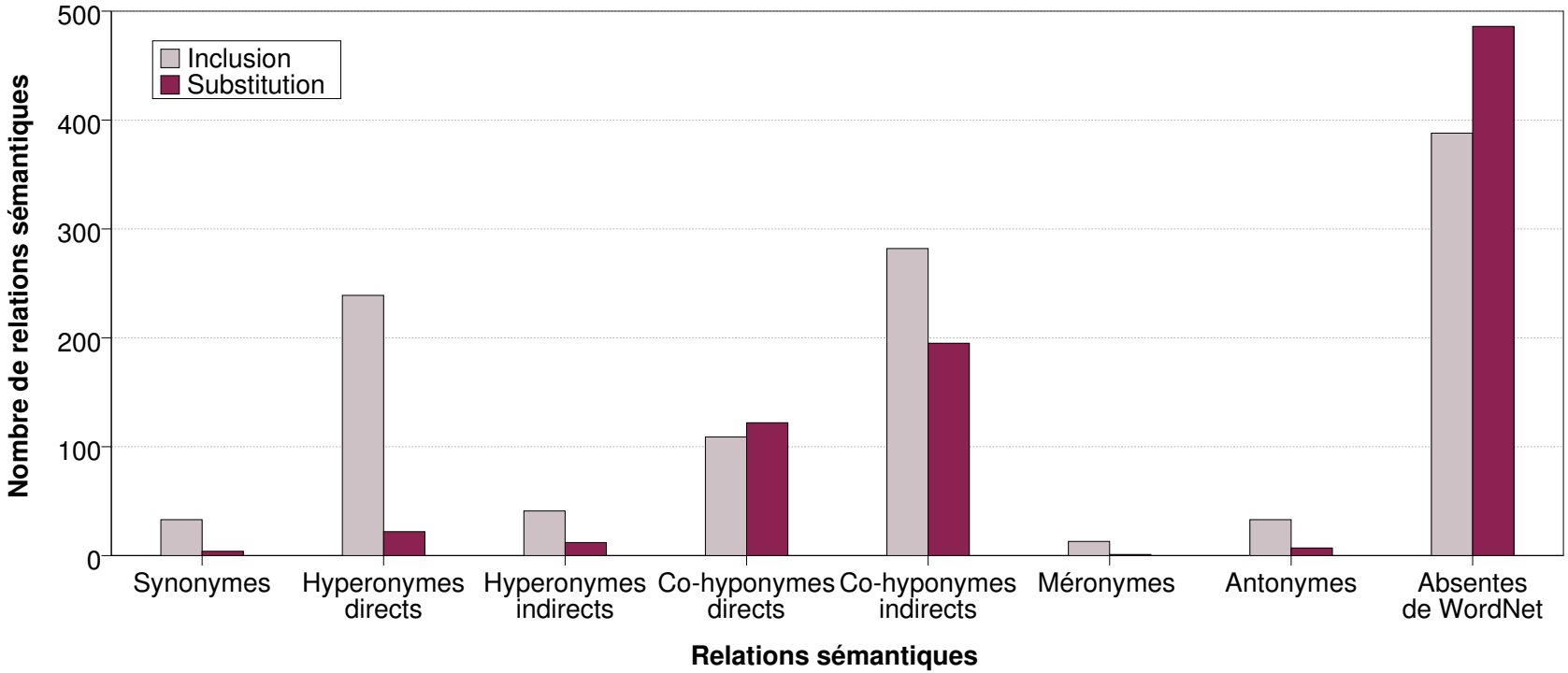
Relations sémantiques identifiées

Comparaison avec le thésaurus du National Cancer Institute (NCIT)



Relations sémantiques identifiées

Comparaison avec WordNet



Analyse des résultats

Synonymie : inclusion

paper, *newspaper*
mistrust, *distrust*

Hyper-/Hyponymie : inclusion

conductor > *semiconductor*

Co-hyponymie : inclusion **et** substitution

hypothalamus et *thalamus* sont co-hyponymes de *neural structure* dans WordNet et co-hyponymes de *Brain_Part* dans le NCIT

Méronymie : préfixes

half-*hour*, *hour*
mid*night*, *night*

Antonymie : préfixes

dis*agreement*, *agreement*
hyper*tension*, ***hypo****tension*

Plan

Conclusion

Apprentissage non supervisé de connaissances morphologiques

Deux approches différentes :

1. Découpage des mots en segments morphémiques
2. Regroupement des mots dans des familles morphologiques

Applications

1. Identification et visualisation des mots clés d'un corpus
2. Acquisition de relations sémantiques

Retour sur les objectifs et les méthodes

- Travail sur corpus
  - Corpus construits automatiquement
  - Données réalistes
- Langue de spécialité
  - Deux thématiques : médecine et sciences de la terre
- Apprentissage et approche statistique
  - Pas de données externes au corpus
- Indépendance aux langues
  - français et l'anglais
  - + finnois, turc et allemand pour le système d'analyse par segmentation

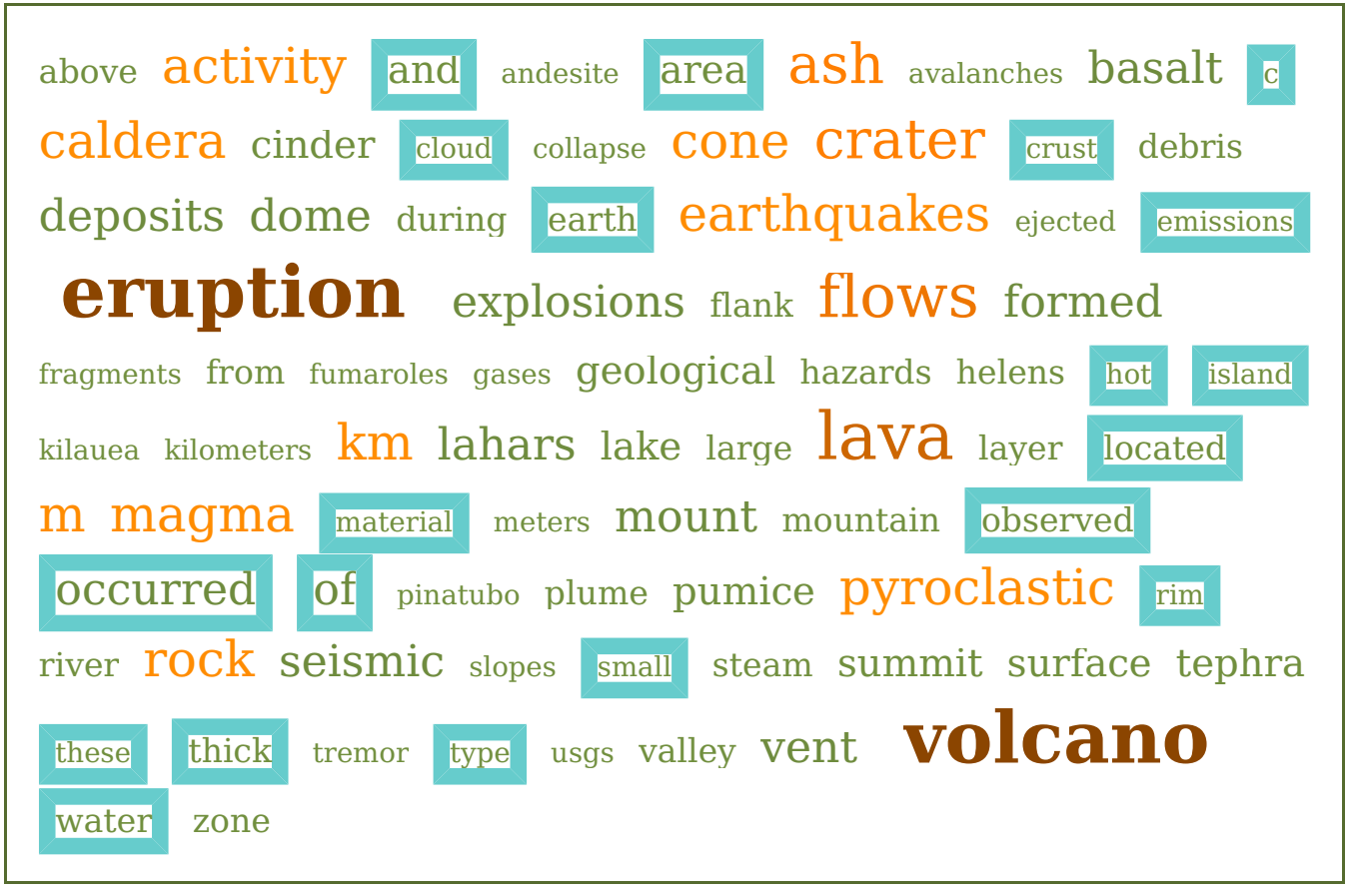
Perspectives

- Amélioration des systèmes d'apprentissage de connaissances morphologiques
- Utilisation des informations contextuelles
- Évaluation pour d'autres applications et d'autres langues
- Morpho Challenge 2007

Merci pour votre attention

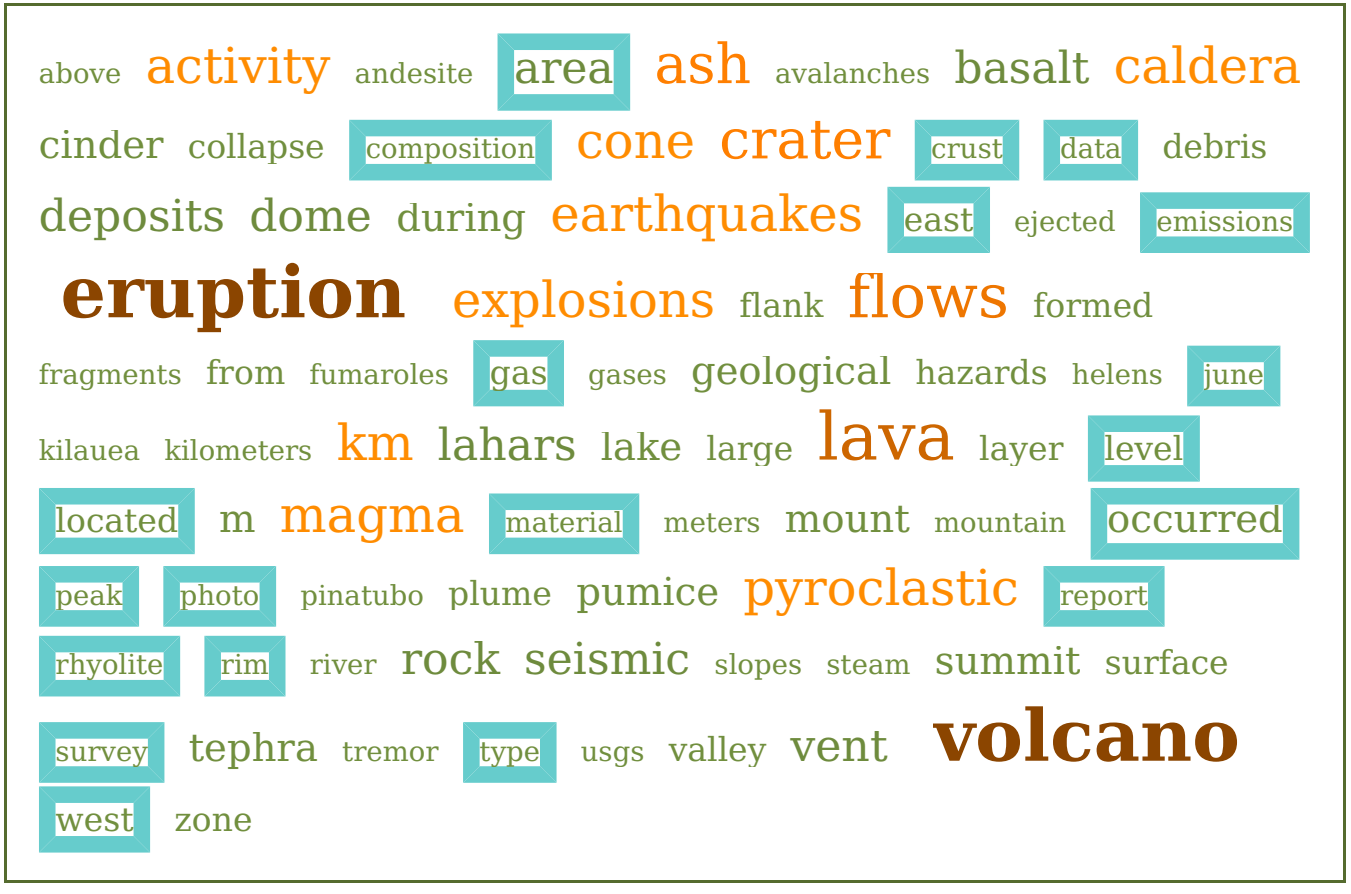
Rôle du corpus de référence [1]

Corpus de référence : collection de corpus de l'université de Leipzig (presse)



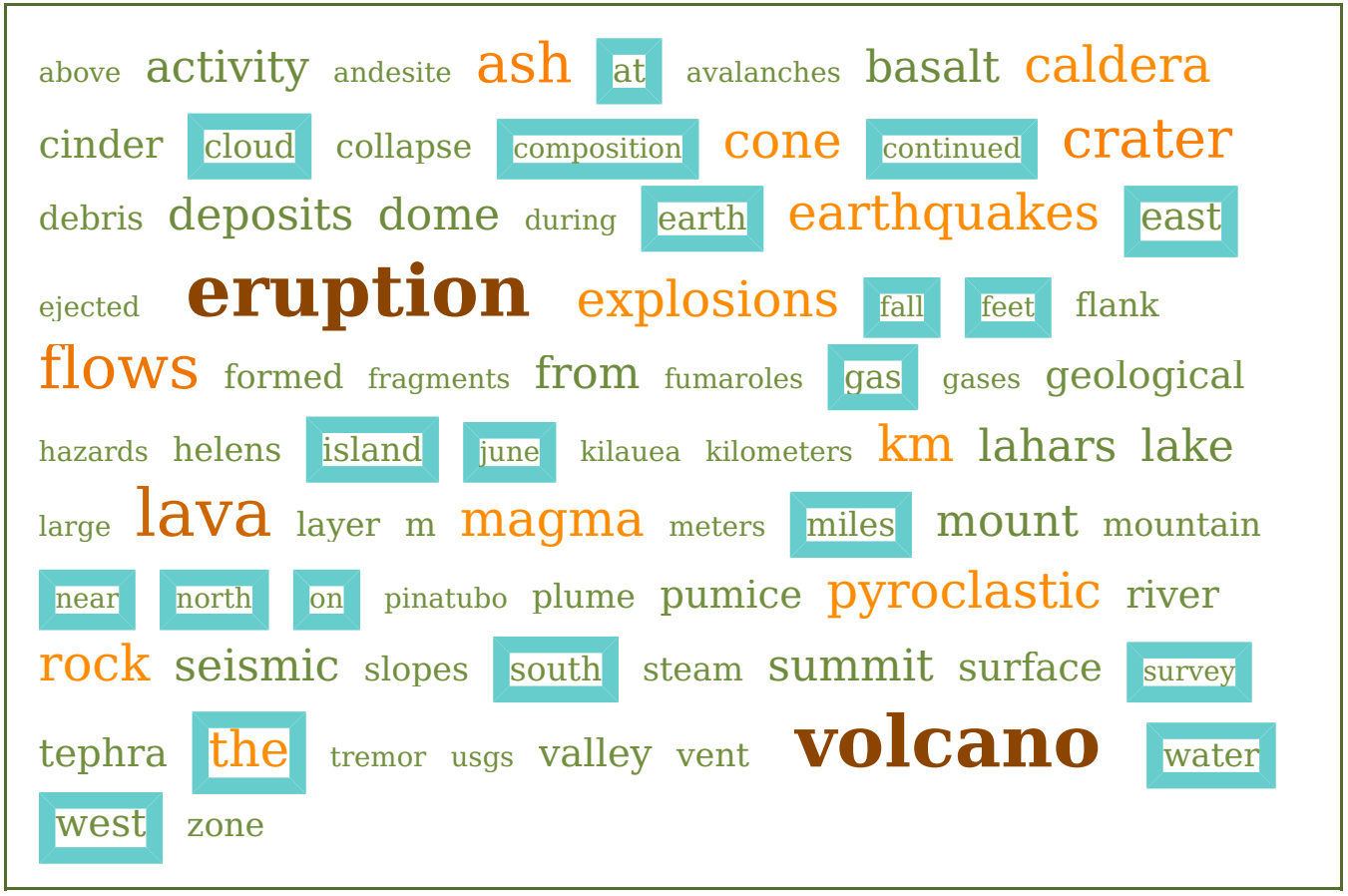
Rôle du corpus de référence [2]

Corpus de référence : liste de mots anglais de Morpho Challenge (projet Gutenberg, corpus Gigaword et Brown)



Rôle du corpus de référence [3]

Corpus de référence : corpus médical



Inclusion et co-hyponymie

Même niveau dans la hiérarchie

- Antonymie
  inactivity, activity
  nonsmoker, smoker
- Unités de mesures :
  kilovolt, volt
  tablespoon, spoon
- Position :
  hypothalamus, thalamus
  parathyroid, thyroid

L'absence de segment morphémique est porteuse de sens