HAL
open science

Analyse automatique de documents botaniques: le projet Biotim

Rousse Guillaume, Éric Villemonte de La Clergerie

▶ **To cite this version:**

Rousse Guillaume, Éric Villemonte de La Clergerie. Analyse automatique de documents botaniques: le projet Biotim. proc. of TIA'05 : Journées "Terminologie & Intelligence Artificielle, Apr 2005, Rouen, France, France. inria-00000119

HAL Id: inria-00000119

https://inria.hal.science/inria-00000119v1

Submitted on 20 Jun 2005

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Analyse automatique de documents botaniques: le projet Biotim

Guillaume Rousse et Éric de la Clergerie

INRIA - Domaine de Voluceau, Rocquencourt BP 105, 78153 Le Chesnay, France
{Guillaume.Rousse,Eric.De_La_Clergerie}@inria.fr

Résumé

À l'heure actuelle, seule une infime partie des connaissances sur la biodiversité sont numérisées. L'écrasante majorité se trouve toujours sous forme de documents papier uniquement, et la saisie manuelle de cette information n'est guère envisageable. Pourtant, la forme hautement structurée de ces documents se prête aisément à l'extraction informatisée. Le projet Biotim essaye justement de s'attaquer à ce problème, en réunissant différentes équipes spécialisées dans l'analyse automatique du texte et de l'image. Cet article présente le travail mené au sein de l'équipe Atoll, concernant l'analyse du langage, et plus particulièrement l'extraction terminologique, en exposant l'originalité du sujet et l'état d'avancement actuel des travaux.

Mots-clés : systématique, botanique, terminologie, ontologie, extraction de connaissances

1. Présentation du projet

1.1. Objectifs

L'objectif du projet Biotim est de concevoir des méthodes génériques d'analyse automatique de masses de données regroupant textes et images pour acquérir une sur-couche sémantique commune et, à partir de ce premier résultat, développer des méthodes génériques d'interrogation pluri-modale des données ainsi structurées.

1.2. Originalité des corpus

Le travail s'effectue sur des corpus botaniques, tout en s'attachant à conserver une démarche générique. Deux ensembles de données sont utilisés. Le premier s'inscrit dans la valorisation de fonds scientifiques botaniques, et concerne des flores d'Afrique de l'Ouest. Le second s'inscrit dans l'étude de données d'expression des gènes suite à une expérimentation à grande échelle, et concerne des descriptions et des photos standardisées de plants d'*Arabidopsis thaliana* mutés.

La flore du Sénégal, sur laquelle s'effectue la majorité du travail d'analyse de texte, est un corpus composé d'une quarantaine de volumes publiés entre 1963 et 2001. Malgré cet étalement dans le temps, la structure des documents reste relativement constante. Comme tout ouvrage de

systématique, il s'agit principalement d'une succession de sections, chacune consacrée à une espèce, ordonnées selon la classification de celles-ci.

Chaque section comprend plusieurs parties à la morphologie bien distinctes. Certaines parties comme la synonymie (c'est-à-dire le fait que l'espèce ait été décrite plusieurs fois sous des noms différents) (Fig. 1) ou l'énumération des spécimens de référence ont une structure très formelle avec des formats hautement condensés et utilisant de nombreuses abréviations. La partie descriptive (Fig. 2) est une énumération des caractères distinctifs de l'espèce, sous forme de phrases nominales juxtaposées. On peut noter une grande quantité d'adjectifs (dénotant essentiellement des formes, couleurs et textures) ainsi que d'adverbes de fréquence et d'intensité (légèrement, particulièrement). De nombreuses entités nommées dénotant des dimensions sont également présentes. Ainsi que des fautes typographiques résultant d'erreurs de reconnaissance de caractères.

Dans le cadre de Biotim, les descriptions sont les parties les plus pertinentes pour l'extraction d'information à l'aide de techniques linguistiques.

= F. brieyi Vermoesen ex G. Gilbert.
Aubréville et Pellegrin, Notul. System. 14,1 : 60 (i95o). - Pellegrin, Bull. Soc. Bot. France 84 : 640 (1987). - Gilbert, Bull. Jard. Bot. État Bruxelles 28,4 : 377 (ig58) ; Fl. Congo belge 7 : 84 (ig58). - Heitz, Forêt du Gabon : 188 et pi. 56 (i943) (sous dénomination « olon » : F. macrophylla var.). - Bois Forêts Tropiques 10 : 175 (1949). -Walker et Sillans, PL utiles Gabon : 38i (1961).

FIG. 1 – Synonymie d'un taxon

Pétiole (o à 3 cm) et rachis portant quelques aiguillons droits, courts (2-3 mm), i coniques et trapus ; rachis légèrement aplati à sa face supérieure et particulièrement au niveau de l'inser- tion des 12-25 paires de folioles. Folioles opposées ou subopposées (même alternes pour les folioles inférieures), sessiles ou subsessiles, de forme oblongue à oblongue-lancéolée, mesurant de 4 X 2 cm pour les folioles inférieures à i5 (-20) X 4 (-5) cm pour les folioles supérieures, les folioles terminales étant toutefois de taille un peu plus réduite que ces dernières et la foliole terminale étant écartée de o à 12 cm de la paire de folioles subterminale ;

FIG. 2 – Description d'un taxon

1.3. Organisation du travail

L'objectif de Biotim est de dégager une méthodologie et des outils pour pouvoir extraire le maximum d'information de corpus descriptifs de taxons, que ce soit en botanique ou, à terme, dans d'autres disciplines similaires. Cette extraction d'information s'effectue en plusieurs étapes, qui ne sont pas encore toutes effectuées :

1. La première étape concerne la correction des artefacts résultants de la numérisation de documents écrits. Parmi ces artefacts, on trouve des erreurs de reconnaissance de caractères, ainsi qu'une structure physique imposée par le support papier (pagination).
2. La seconde étape consiste à distinguer les différentes parties des corpus, de façon à pouvoir appliquer des traitement différenciés en fonction de la position dans celui-ci, c'est-à-dire tirer parti du contexte pour améliorer l'analyse.

3. La troisième étape comporte une phase d'analyse morpho-syntaxique des corpus. Un pipeline formé de plusieurs outils existant (segmenteur, reconnaisseur d'entités nommées, étiqueteur, extracteur de terminologie) permet de combiner et de comparer plusieurs résultats, de façon à fournir une analyse plus robuste. En sortie de ce pipeline, il est possible d'exploiter des outils d'extraction terminologique pour effectuer des analyses statistiques, comme la comparaison des ensembles gouverneur-gouvernés pour chaque terme.

4. La quatrième étape comporte une phase d'analyse syntaxique des corpus pour identifier des dépendances syntaxiques entre mots. L'examen de ces dépendances en utilisant des techniques d'apprentissage doit permettre l'extraction semi-automatique d'une ontologie d'un domaine, en s'appuyant sur l'hypothèse distributionelle de Harris (Harris, 1968) que des mots partageant des contextes syntaxiques similaires sont sémantiquement proches. L'utilisation de marqueurs syntaxiques plus spécifiques est également envisageable : énumérations, intervalles (comme « de forme oblongue à oblongue-lancéolée »), ... L'ontologie résultante doit ainsi pouvoir indiquer que « sessile » est un adjectif de structure s'appliquant aux « folioles » (eux même organes de plantes).

5. Cette dernière étape consiste à extraire pour chaque taxon l'ensemble des propriétés décrites. Cette étape s'appuie sur l'analyse syntaxique des corpus en utilisant l'ontologie précédemment extraite pour lever un maximum d'ambiguïtés.

2. Situation actuelle

2.1. Préparation des corpus

Le début du travail a été marqué par de lourdes difficultés, dues à la mauvaise qualité du travail de numérisation. Nous les évoquons ici pour signaler que, d'une certaine manière, elles font partie intégrante de toute tentative de valorisation de fonds scientifiques anciens qui commence par une telle phase de numérisation.

Les documents numérisées que nous avons obtenus présentaient un très fort taux d'erreurs typographiques. Diverses tentatives de corrections automatiques ont été expérimentées pour y pallier. L'idée générale consiste à rechercher par analyse statistique les mots considérés comme erronés, puis à utiliser des systèmes à automates finis pour trouver des approximations dans un lexique. Néanmoins, en présence d'un vocabulaire très spécifique pas forcément présent dans des lexiques, nous avons complété ce lexique en recherchant les mots inconnus dans les corpus ayant une fréquence relativement forte. Malheureusement, le manque de document de référence (corrigés) rend difficile la validation des différentes heuristiques utilisées.

Il est ensuite apparu que certains corpus avaient été numérisés en double page de telle manière que les paragraphes étaient alternés entre pages gauches et droites. Enfin, la numérisation était orientée vers un rendu visuel proche de l'original, afin de permettre une consultation identique sur support numérique, et absolument pas vers une extraction de texte comme objectif principal.

Au final, il a été décidé une renumérisation intégrale des corpus, selon une méthodologie homogène et adaptée aux besoins. Nous venons de recevoir les corpus renumérisés. Il reste néanmoins certaines erreurs, comme le montrent les extraits précédents (Fig. 1 et Fig. 2).

2.2. Traitement morpho-syntaxique

L'analyse morpho-syntaxique, une fois les étapes préliminaires effectuées, s'appuie sur une chaîne de traitement intégrant différents outils. L'objectif de cette chaîne est d'être ouverte et de pouvoir intégrer facilement de nouveaux composants.

Pour cela, nous nous appuyons sur une représentation XML (Listing 3) inspirée par la proposition de standardisation des annotations morpho-syntaxiques (MAF (Clément & Villemonte de la Clergerie, 2004)). Cette proposition identifie un niveau associé à la segmentation et un autre associé aux unités linguistiques ou mots, généralement associées à des entrées lexicales et portant l'information morpho-syntaxique.

Les composants ajoutés sont pris en charge par un programme principal dont le rôle est d'extraire certaines informations du flux XML (en s'appuyant sur des requêtes de type XPath), de les envoyer au composant proprement dit et de récupérer les données retournées par celui-ci pour les incorporer au flux initial.

Le flux XML peut représenter un ensemble d'alternatives exprimées comme les transitions d'un automate à état fini et les informations morpho-syntaxiques sont exprimables soit sous forme compacte d'étiquettes ou sous la forme expansée de structures de traits.

Le premier outil intégré dans la chaîne est le segmenteur TOKENIZER, développé au sein de l'équipe Atoll. Il se fonde sur un lexique pour, d'une part, effectuer son découpage, et, d'autre part, pour récupérer les informations associées aux mots connus. On obtient donc une liste de tokens, auxquels sont associés un ou plusieurs mots en cas d'ambiguïté d'analyse, chaque mot pouvant correspondre à une ou plusieurs entrées du lexique.

L'étiqueteur TREETAGGER [1] est également appelé sur les tokens (niveau segmentation) pour proposer une information morpho-syntaxique concurrente de celle proposée par le lexique. Il est utilisé ici après ré-entrainement sur un corpus propre à l'équipe, de façon à fournir des réponses plus fines.

Le lemmatiseur FLEMM (Namer, 2000) peut éventuellement être appelé sur les mots inconnus mais cette phase est actuellement désactivée.

L'agrégation des informations fournies par les divers outils nécessite une normalisation préalable, chaque outil utilisant un formalisme différent pour exprimer différentes notions comme :
- le doute : certains outils donnent un résultat par valeur possible, certains donnent un résultat combiné ;
- les entités nommées ;
- les informations morphologiques elles-même.

Les deux premières notions sont simples à résoudre par le biais du programme encapsulant l'appel à ces outils, car il existe une relation directe et non-ambiguë entre chacune des notations utilisées. Par exemple, TREETAGGER utilise « <unknown> », tandis que TOKENIZER utilise « -unknown- », mais dans les deux cas la notion sous-jacente est la même.

Le dernier point est par contre beaucoup plus délicat. Tout d'abord parce que les notations sont complexes. Par exemple, la notation dérivée de Multext utilisée dans notre lexique utilise « A-qual-fs » pour désigner un adjectif qualificatif féminin singulier. Mais surtout parce qu'il n'existe pas forcément une bijection entre les différents espaces de valeurs de chaque outil. Multext différencie ainsi plusieurs catégories d'adjectifs, là où TREETAGGER n'en connaît

[1] `http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/`

qu'une seule. Au contraire, la notion de cardinal pour TREETAGGER correspond selon les cas à un adjectif cardinal, un déterminant ou à un nom pour Multext.

Un outil de conversion à vocation universelle a donc été développé. Il se fonde sur un arbre de décision pour convertir une valeur d'entrée dans un format pivot en structure de traits, qui est a son tour converti selon le même principe dans une valeur de sortie. Cet outil connaît à ce jour 4 dialectes : Multext, TreeTagger, Fastr, Acabit. Ajouter un dialecte supplémentaire revient à ajouter un convertisseur vers le format pivot, ce qui le rend facilement extensible. Il est disponible sous forme de module Perl sur CPAN [2].

Une fois ces différents résultats agrégés, une série d'heuristiques permet de rendre un résultat non ambigu :

- les mots composés sont préférés aux mots simples ;
- le choix du segmenteur, fondé sur un lexique, est prioritaire, sauf dans le cas des entités nommées comme les noms propres ;
- en cas de doute (plusieurs entrées possibles du lexique) ou de mot inconnu, le choix de l'étiqueteur est pris en compte.

```
<token id="t20724" value="rachis">rachis</token>
<wordForm entry="rachis" lemma="rachis" tag="cat@noun type@common
   gender@masc" tokens="t20724"/>
<token id="t20725" value="portant">portant</token>
<wordForm entry="portant" lemma="porter" tag="cat@verb mode@part
   tense@pres" tokens="t20725"/>
<token id="t20726" value="quelques">quelques</token>
<wordForm entry="quelques" lemma="quelques" tag="cat@det num@pl"
   tokens="t20726"/>
<token id="t20727" value="aiguillons">aiguillons</token>
<wordForm entry="aiguillons" lemma="aiguillon" tag="cat@noun
   type@common gender@masc num@pl" tokens="t20727"/>
```

FIG. 3 – Annotations morpho-syntaxique

2.3. Extraction terminologique

À l'issue de la chaîne de traitement morpho-syntaxique, il est possible et intéressant d'appliquer des outils d'extraction terminologique comme FASTR (Jacquemin *et al.*, 1997) ou ACABIT (Daille, 2003). Cette extraction terminologique a diverses motivations. D'une part, elle doit permettre de produire un lexique du domaine de manière à faciliter l'analyse syntaxique ultérieure. D'autre part, la terminologie produite est utile à terme, comme ressource documentaire, pour indexer ou rechercher des documents dans le cadre d'un système d'information botanique. Enfin, nous utilisons cette terminologie pour des expériences préliminaires d'acquisition de classes sémantiques.

L'application des outils d'extraction terminologique est facilité par les outils de conversions entre jeux d'étiquettes. Nous avons ainsi pu mener des expériences avec ACABIT, que nous souhaitons maintenant répéter avec FASTR.

Les corpus étudiés ne se prêtent pas forcement très bien à l'extraction terminologique telle qu'effectuée par un outil comme ACABIT. En effet, il n'y a finalement pas énormément de

[2] http://search.cpan.org/~grousse/Lingua-TagSet

arbuste lianescent	arbuste iriermes	arbuste à feuille
arbuste inerme	arbuste épine	port d'arbuste
arbuste épineux	arbuste ornemental	premier arbuste
arbuste sarmenteux	arbuste lianescents	massif d'arbuste
arbuste de sous-bois	arbuste à jeune	espèce d'arbuste
arbuste à coloniser	arbuste à rameau	...

FIG. 4 – Vocabulaire extrait

groupes nominaux complexes mais plutôt un usage très intensif d'adjectifs modifiants des noms et ce à très longue distance. Néanmoins, l'outil rend beaucoup d'information et il est difficile, sans experts, de déterminer la validité de la terminologie ainsi extraite.

2.4. de la terminologie vers des classes sémantiques

Pour faciliter la compréhension et la validation de terminologie, nous explorons diverses méthodes de visualisation graphiques sous forme de cartes ou de réseaux d'associations.

À chaque terme est associé l'ensemble des autres termes auxquels il est lié par une relation de gouverneur, et l'ensemble de ceux auxquels il est lié par une relation de gouverné. Ces deux ensembles, utilisés simultanément ou non selon l'analyse, constituent le contexte de ce terme.

La figure 5(a) montre une vue globale d'un tel réseau d'associations, sur un seul volume du corpus, et avec un niveau de filtrage fondé sur la productivité des contextes utilisés. Il illustre la concentration du vocabulaire autour de termes pivots, qui apparaissent ici comme des noeuds fortement ramifiés. La figure 5(b) montre une vue locale de ce même graphe, centré autour d'un tel terme. « Teinte » apparaît d'une part comme le gouverneur des termes « jaune », « vert », « bleu », « pourpre », « noir », « rouge », « rosé » et « clair », et d'autre part comme le gouverné des termes « feuille », « ovaire », « inflorescence », « aiguillon » et « rachis ». On remarque une certaine homogénéité de ces ensembles, avec d'une part un ensemble de couleurs, et d'autre part, un ensemble d'organes.

D'une façon plus générale, le lien gouverneur-gouverné implicitement présent dans une entrée terminologique peut être exploité pour tenter des regroupements sémantiques, sous l'hypothèse distributionnelle (Harris, 1968) que des mots liés à des ensembles similaires sont sémantiquement proches. Un coefficient de similarité peut être établi entre chaque terme, sur la base de la comparaison de leur contexte. Quatre coefficients tirés de la littérature ont été retenus à l'issue d'un travail antérieur (Pochon, 2003), mais nous manquons néanmoins d'éléments de validation pour les comparer entre eux.

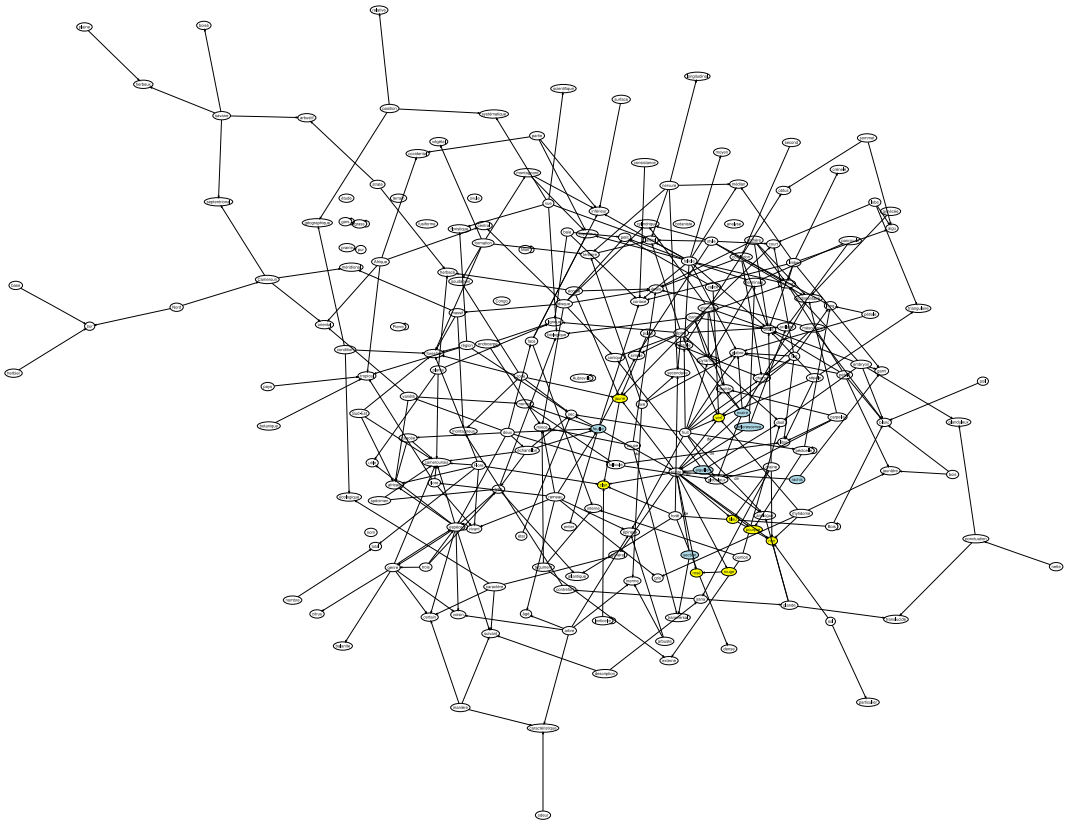
– Le coefficient *a* (Denoue & Vignollet, 1996) correspond au nombre d'éléments communs entre deux contextes.

$$a(t_1, t_2) = |c_1 \cap c_2| \tag{1}$$

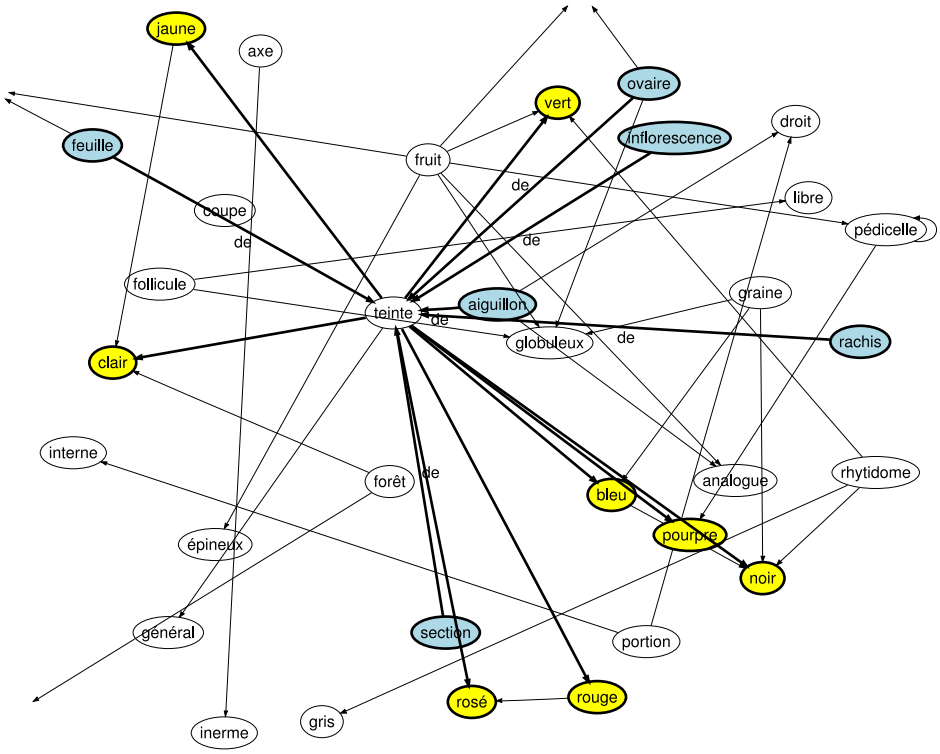
– Le coefficient *prox* (Denoue & Vignollet, 1996) correspond à la somme pour chaque élément commun de l'inverse de leur productivité élevée au carré, la productivité se définissant comme le nombre de contextes dans lesquels ils apparaissent.

$$prox(t_1, t_2) = \sum \frac{1}{prod(c)^{1/2}} \tag{2}$$

– Le coefficient *j* (Denoue & Vignollet, 1996) correspond au produit du nombre d'éléments partagés par la taille du contexte, calculé pour chaque terme et moyenné pour obtenir une

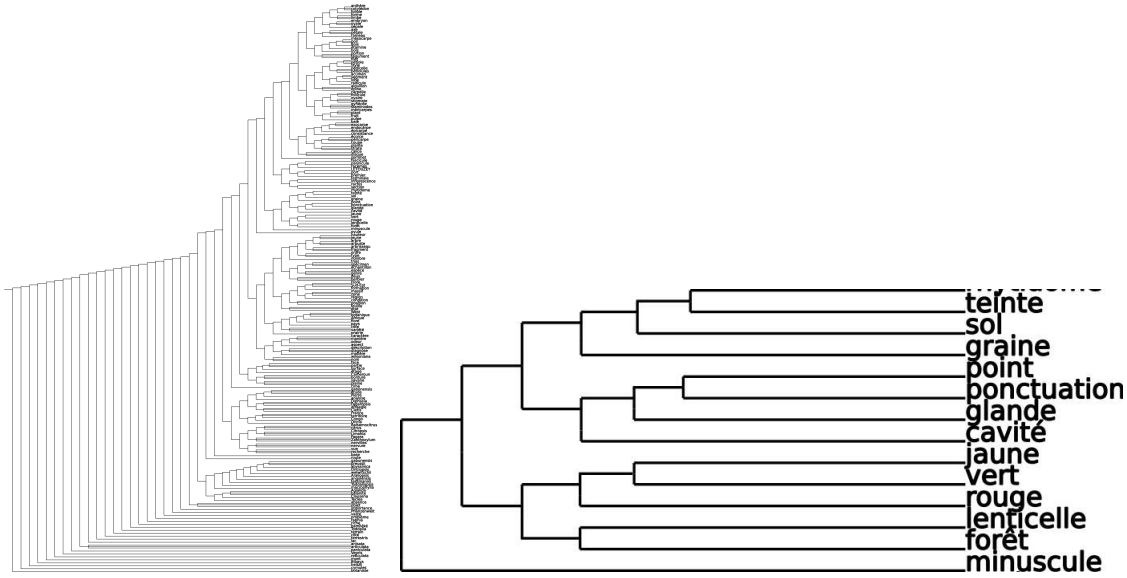


(a) vue globale



(b) détail

FIG. 5 – Relations entre termes



(a) vue globale (b) détail

FIG. 6 – Distance entre termes

similarité symétrique.

$$j(t_1, t_2) = (\frac{|c_1 \cap c_2|}{|c_1|} + \frac{|c_1 \cap c_2|}{|c_2|})/2 \quad (3)$$

– Le coefficient *dice* (Bourigault & Lame, 2002) correspond au produit du nombre d'éléments partagés par la moyenne de la taille des contextes.

$$dice(t_1, t_2) = \frac{|c_1 \cap c_2|}{(|c_1| + |c_2|)/2} \quad (4)$$

Une fois la matrice des distances obtenue à partir d'un de ces coefficients, il est possible d'utiliser un algorithme de classification ascendante hiérarchique pour obtenir un arbre. La figure 6(a) présente une vue générale d'un tel arbre, avec le même niveau de filtrage que pour le graphe précédent, et utilisant le coefficient *dice*. La figure 6(b) présente une vue de détail du même arbre, centrée autour d'une partie des termes gravitant autour de « teinte ». Les couleurs sont effectivement regroupées ensemble, mais pas les termes anatomiques, qui sont dispersés ailleurs dans l'arbre.

Une autre piste encore inexplorée consiste à tirer partie de la force de ces liens telle que fournie par ACABIT pour orienter l'algorithme de regroupement.

2.5. Vers des dépendances syntaxiques

Cette expérience de regroupement sémantique préfigure le processus d'extraction d'ontologies que nous devons prochainement mener. Au lieu de considérer des liens très généraux gouverneur-gouverné entre mots proches dans la phrase, nous comptons nous appuyer sur des liens de dépendance issus de l'analyse syntaxique. Cela doit nous permettre d'accéder à une caractérisation plus fine des dépendances entre mots plus éloignés. Ce programme est très similaire à celui proposé dans (Bourigault, 2002). Les différences portent sur la nature de l'analyse

syntaxique qui, dans notre cas, retourne des analyses profondes et ambiguës. Nous comptons donc nous appuyer sur des informations riches, ambiguës et éventuellement plus précises pour mener la phase d'apprentissage.

Pour cela, nous allons utiliser un analyseur syntaxique compilé avec le système DYALOG (Villemonte de la Clergerie, 2002) à partir d'une grammaire d'arbres adjoint (TAG) et à large couverture, elle-même produite à partir d'une méta-grammaire du français. Cet analyseur syntaxique a été testé à grande échelle dans le cadre de la campagne d'évaluation syntaxique EASY sur plus de 30000 phrases. Son taux actuel de couverture pour des analyses de phrases complètes est de quasiment 100% sur un jeu de 321 phrases tests issues d'EUROTRA et de 93% sur un jeu de 1661 phrases issues de TSNLP. L'analyseur peut également retourner des analyses partielles si aucune analyse complète n'est trouvée et sait gérer la présence de mots inconnus.

D'autre part, l'utilisation d'une méta-grammaire permet une description linguistique modulaire sous forme d'un ensemble de classes organisées dans un réseau d'héritage, chaque classe fournissant des éléments d'information sur un phénomène syntaxique. Cette organisation modulaire permet de relativement facilement désactiver les classes associées à des phénomènes linguistiques peu ou pas présents dans les corpus étudiés ou au contraire d'ajouter de nouvelles classes (voire de compléter certaines classes existantes) pour certains phénomènes linguistiques spécifiques aux corpus. Ce travail d'ajustement permettra d'améliorer la couverture tout en réduisant l'ambiguïté des analyses.

En effet, à ce stade, faute d'informations complémentaires (à savoir celles que nous cherchons justement à obtenir), il n'est pas en général possible de produire des analyses non ambiguës, en particulier au niveau des rattachements prépositionnels, voire de certaines coordinations. L'ensemble de toutes les analyses peut cependant être produit et représenté par l'analyseur sous la forme compacte d'une forêt de dérivations. Ces forêts sont convertibles en forêts partagées de dépendances entre têtes des constituants, comme illustré par la figure 7 pour « des sépales ovales-aigues, glabres ou éparsement hérissés ». Nous envisageons d'appliquer un algorithme de filtrage et conversion sur ces forêts de dépendances pour extraire et normaliser les dépendances les plus pertinentes. Si réellement nécessaire, cet algorithme, qui devrait s'appuyer sur un algorithme similaire développé pour la campagne EASY, peut également effectuer un travail préliminaire de désambiguisation, fondé sur quelques heuristiques.

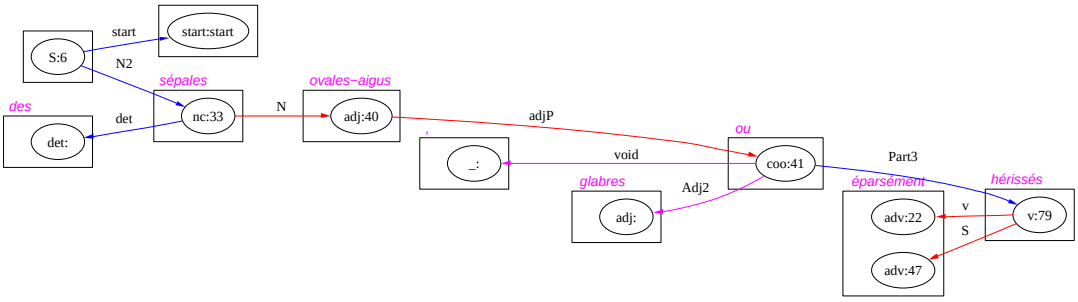


FIG. 7 – Forêt de dépendances

L'utilisation de techniques d'apprentissage doit ensuite permettre de repérer les dépendances les plus fortes (celles qui reviennent fréquemment et avec le moins d'ambiguïté) servant ensuite à faire émerger des classes sémantiques et des relations entre ces classes. Par itération et réduction progressive d'ambiguïtés, d'autres dépendances seront alors renforcées.

La présence de certaines constructions syntaxiques type énumérations ou intervalle doit éga-

lement permettre d'assurer un étiquetage sémantique de certaines classes en partant d'une ontologie germe. Certaines constructions explicite comme « de forme X » ou « de couleur Y » seront également utilisées. Le fait de savoir qu'un mot désigne une couleur permet ensuite par percolation de transmettre cet étiquetage sémantique.

3. Conclusion

Les premières expériences déjà menées au niveau terminologique préfigurent celles devant s'effectuer au niveau syntaxique. L'ensemble des éléments nécessaires à l'extraction de l'ontologie devraient être opérationnel dans les mois à venir. Les résultats déjà obtenus sont relativement prometteurs mais soulignent l'importance du travail de validation en relation avec des experts en botanique. En effet, nous nous contentons pour le moment d'explorer des pistes diverses, sans possibilités de décider entre celles-ci en l'absence de procédure d'évaluation. Ces experts étant en général très occupés et pas nécessairement familiers avec les aspects langagiers, cette étape passe forcément par la recherche de méthodes de visualisation leur facilitant la tâche.

Références

BOURIGAULT D. (2002). UPERY : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. In *Proc. of TALN'02*, p. 75–84, Nancy, France.

BOURIGAULT D. & LAME G. (2002). Analyse distributionnelle et structuration de terminologie. application à la construction d'une ontologie documentaire du droit. *TAL*, **43**, 129–150.

CLÉMENT L. & VILLEMONTE DE LA CLERGERIE E. (2004). Terminology and other language resources – morpho-syntactic annotation framework (MAF). ISO TC37SC4 WG2 Working Draft.

DAILLE B. (2003). Terminology mining. In M. P. (ED), Ed., *Information Extraction in the Web Era*, Lectures Notes in Artifi al Intelligence, p. 29–44. Springer.

DENOUE L. & VIGNOLLET L. (1996). L'importance des annotations. application à la classifi cation des documents du web. *L'indexation conceptuelle et structurelle*, **1**, 1–22.

HARRIS Z. (1968). *Mathematical Structures of Languages*. New-York : John Wiley & Sons.

JACQUEMIN C., KLAVANS J. L. & TZOUKERMANN E. (1997). Expansion of multi-word terms for indexing and retrieval using morphology and synta. In *Proceedings of ACL-EACL'97*, p. 24–31, Madrid, Spain.

NAMER F. (2000). Flemm : Un analyseur flexionnel du français à base de règles. *Traitement automatique des langues pour la recherche d'information, revue T.A.L.*

POCHON A. (2003). Intégration d'une méthode d'acquisition de terminologie et recherche de relations. Mémoire de DEA, LIFO – Université d'Orléans.

VILLEMONTE DE LA CLERGERIE E. (2002). Construire des analyseurs avec DyALog. In *Proc. of TALN'02*.