

**De l'APD à Tropes : comment un outil d'analyse de contenu
peut évoluer en logiciel de classification sémantique généraliste**

Pierre Molette
pierre.molette@acetic.fr

Communication au colloque Psychologie et communication
Tarbes – Juin 2009

Introduction

Afin de s'affranchir des biais de l'analyse thématique (en particulier la définition arbitraire des unités de codage et du choix subjectif des indicateurs utilisés pour l'interprétation), Rodolphe Ghiglione et le Groupe de Recherche sur la Parole (GRP, Université Paris 8) ont élaboré successivement deux théories d'analyse de contenu : l'Analyse Propositionnelle du Discours (APD) puis l'Analyse Cognitive Discursive (ACD).

Ces théories sont fondées sur un découpage du texte en propositions grammaticales, la catégorisation sémantique des mots outils, l'identification de classes paradigmatisées de substantifs (appelés "référents noyaux") et la modélisation des propositions sous un formalisme simplifié (appelé "modèle argumentatif" dans l'APD ou "noyau générateur" et "structure fondamentale de la signification" dans l'ACD). Pour résumer, disons qu'il s'agit d'extraire du texte une série de variables qui vont faire l'objet d'un traitement statistique permettant de révéler des résultats d'analyse objectifs, qui ne seraient pas forcément identifiés après une lecture approfondie du texte.

L'informatisation de la théorie de l'APD a fait l'objet de plusieurs prototypes de logiciels universitaires (dans les années 80 et au début des années 90), qui ont été utilisés pour des travaux de recherche avec des résultats satisfaisants. Ces prototypes, rudimentaires, nécessitaient la présence permanente de l'utilisateur, impliquaient des interventions fréquentes sur le lexique (de capacité insuffisante) et restituaient les résultats sous la forme de tableaux statistiques. L'utilisation était laborieuse : l'analyse durait des heures et il fallait la recommencer pour appliquer une modification de dictionnaire. Il s'agissait en fait d'une "analyse manuelle assistée par ordinateur". On arrivait donc à un paradoxe : la théorie et l'outil existaient, mais ils étaient très lents et difficilement exploitables. Ce qui impliquait, faute de temps, de réduire la quantité de texte traitée et paradoxalement de prendre des risques sur l'interprétation des résultats, puisque ceux-ci étaient issus de petits échantillons de données textuelles.

On s'est donc mis à rêver d'un logiciel automatique, rapide, doté d'une interface graphique moderne. Mais pour y parvenir il fallait se doter de moyens qui dépassaient largement le budget d'un laboratoire universitaire. Le Groupe de Recherche sur la Parole s'est donc appuyé sur une entreprise, qui s'est lancée dans un processus d'ingénierie et a développé le logiciel Tropes, dès 1994. Après sept versions consécutives, et quinze ans d'existence, Tropes est devenu une plate-forme d'analyse sémantique de contenu. Il s'est éloigné des "pures" méthodologies d'origine, qui se sont en quelque sorte "diluées" dans d'autres. L'objectif de cette conférence est de montrer comment ces théories et ces outils issus de la recherche universitaire ont du être transformés, puis fusionnés, pour obtenir une suite d'outils d'analyse sémantiques qui dépassent largement le cadre méthodologique fixé à l'origine par l'APD et l'ACD.

Analyse lexicométrique versus analyse sémantique

La lexicométrie consiste, pour résumer, à trier des formes fléchies extraites d'un corpus de textes, à filtrer les « mots outils » et à supprimer les termes à faible fréquence (dont les hapax), afin d'élaborer des statistiques.

Cette approche est parfaitement valable et utile.

Cependant la lexicométrie pose plusieurs problèmes quand on veut s'en servir pour l'analyse de contenu :

- 1 – L'absence de résolution des ambiguïtés nécessite d'écarter certains termes, qui sont pourtant essentiels.
- 2 – Le traitement statistique implique de réduire fortement le nombre de variables, donc de perdre de l'information.
- 3 – L'analyste doit découvrir lui-même les équivalents sémantiques durant la phase d'interprétation des résultats.
- 4 – Les mots composés ne sont pas reconnus, ce qui introduit de nombreux artefacts.

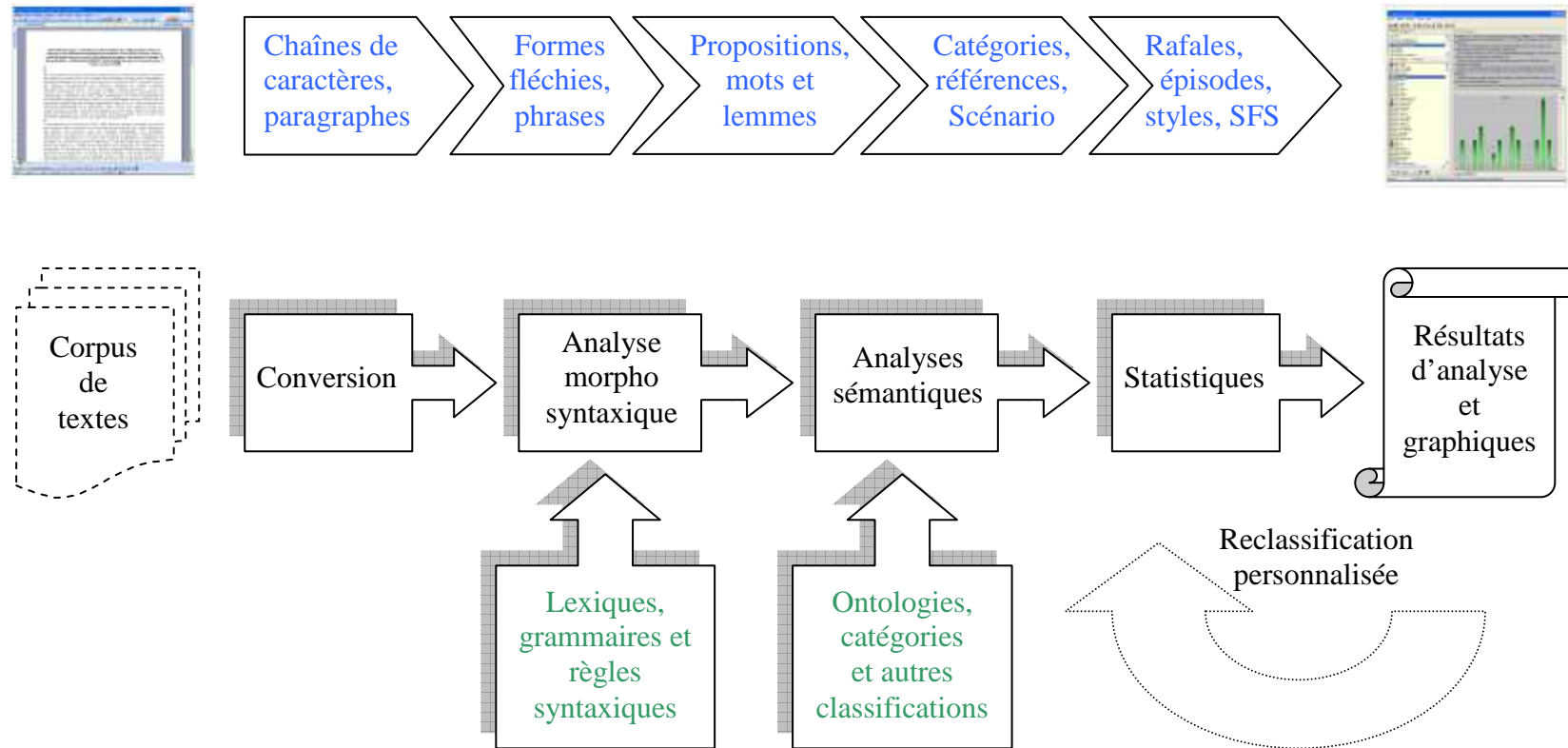
Malgré ces inconvénients, la lexicométrie présente l'intérêt d'être simple du point de vue informatique et de pouvoir fonctionner sans intervention humaine (on la retrouve dans certains moteurs de recherche). Ce qui lui permet de traiter de nombreuses langues vivantes et explique qu'elle soit complétée par de nombreux outils statistiques, plus astucieux les uns que les autres.

A contrario, l'analyse sémantique va regrouper des mots issus d'un corpus de textes dans des catégories (par ex. cause, but, temps, lieu, etc.) ou des classifications (synonymes, hyperonymes), en s'appuyant sur des grammaires et des réseaux sémantiques. Disons qu'on passe de "l'analyse de contenant" (formes) à l'analyse de contenu (sens), en faisant appel à la pragmatique linguistique (i.e. tenir compte du contexte). La sémantique facilite l'analyse et réduit le risque interprétatif.

Mais cette approche implique d'affronter la polysémie, avec l'inconvénient d'utiliser une logique complexe de résolution de problèmes (grammaticaux, sémantiques) et de nécessiter de gros dictionnaires de classification, qui ne seront jamais totalement parfaits ou exhaustifs. L'analyse sémantique impose donc des reclassifications, avant l'interprétation des résultats.

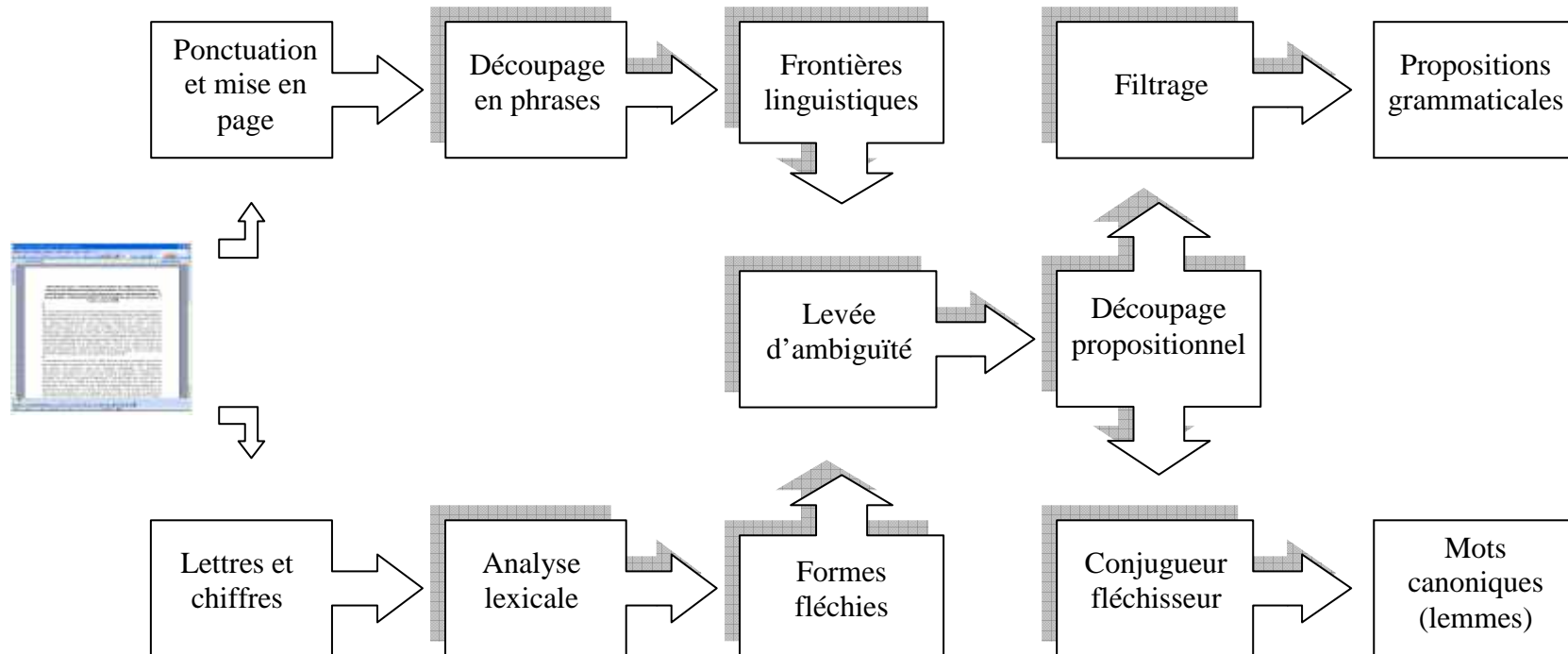
Un exemple : dans un texte qui contient "de l'or, de l'argent, du bronze", trois substantifs qui ne seraient comptés qu'une fois en lexicométrie (donc non significatifs dans ce cadre), Tropes va retenir la classification "métaux et alliages" (comptée trois fois), après désambiguïsation ("or"=>conjonction et métal ; "bronze"=>verbe bronzer, objet d'art et métal ; "argent"=>moyen de paiement et métal) ; en écartant des ambiguïtés comme "livre d'or", "or noir", "médaille d'argent", "âge du bronze", etc.

Tropes - Vue globale du processus d'analyse



Contrairement aux logiciels de lexicométrie, Tropes fait appel à deux processus d'analyse (morphosyntaxique et sémantique) avant de faire des statistiques. Le filtrage des hapax est optionnel.

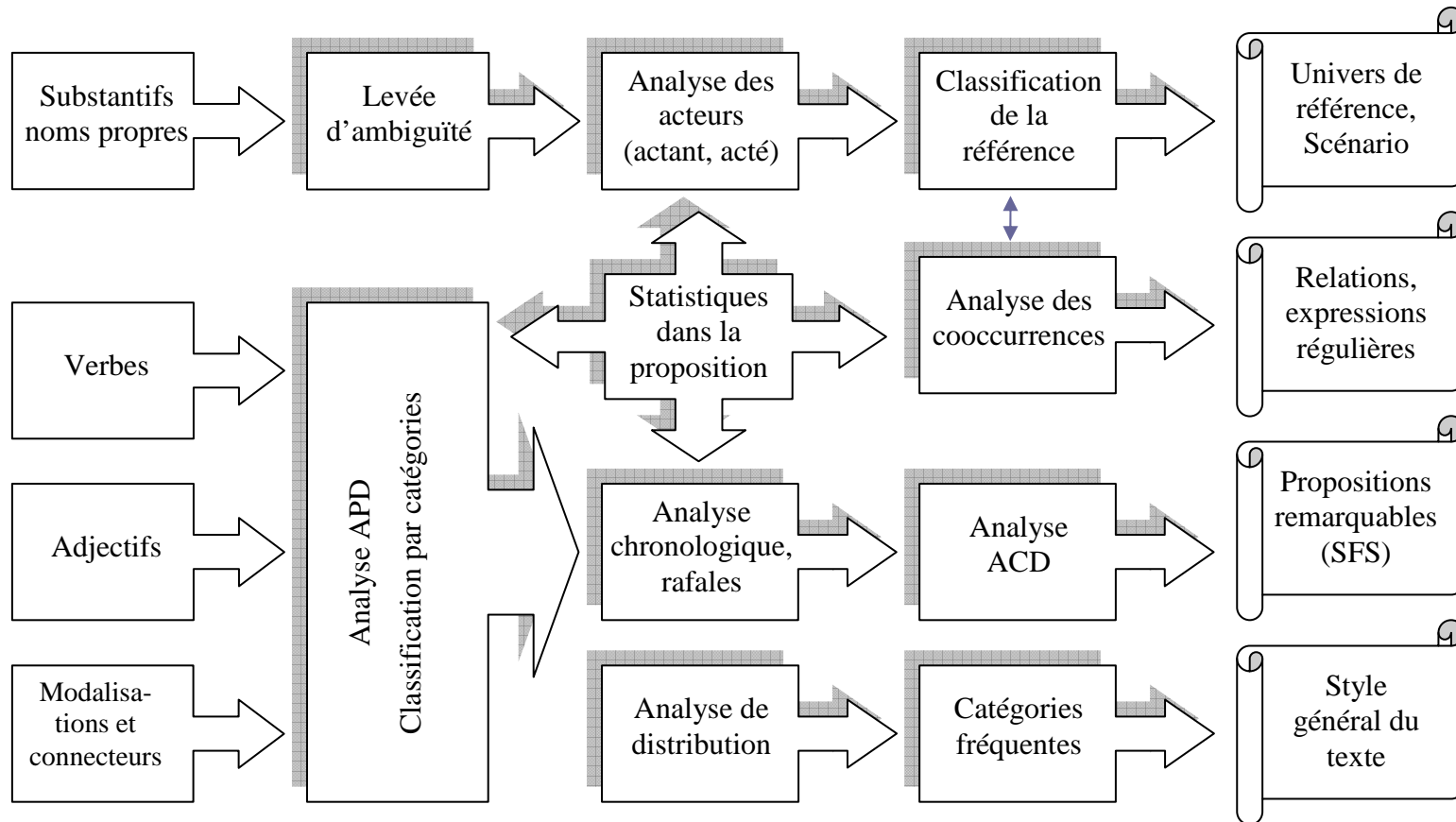
Tropes - Vue globale de l'analyse morphosyntaxique



L'assemblage des mots composés intervient à plusieurs endroits dans ce schéma (avant et après découpage propositionnel).

La levée d'ambiguïté est un processus très complexe, qui fait appel à la logique des prédicats, des grammaires statistiques, un correcteur orthographique, un fléchisseur et un lexique.

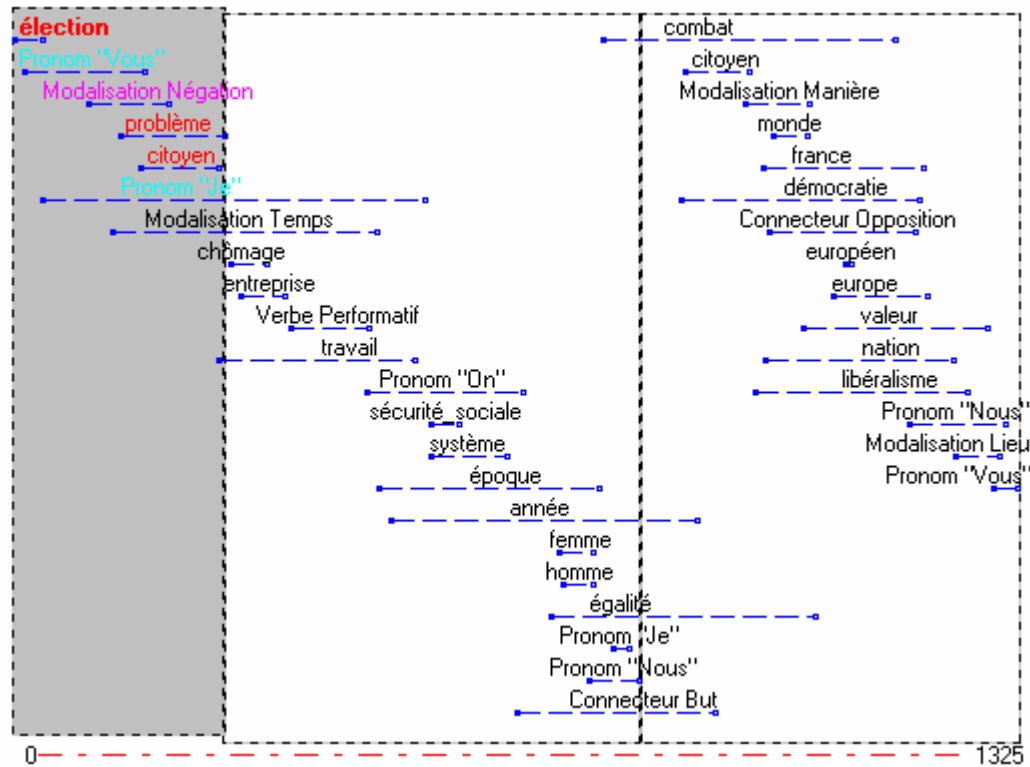
Tropes - Vue globale des analyses sémantiques



La levée d'ambiguïté sémantique fait appel à un processus d'analyse stochastique exploitant une "métaphore informatique" de la mémoire humaine (i.e. un processus de décision incluant des connaissances a priori, une mémoire à court terme et à long terme).

Comment présenter certains résultats ? Des graphiques conçus comme des outils d'analyse

Le graphe des rafales et des épisodes est une analyse chronologique du récit :



Une Rafale regroupe des occurrences (contenues dans une classe d'équivalents ou une catégorie APD) ayant tendance à arriver avec une concentration significative dans une partie limitée du texte (mais jamais de façon uniforme sur l'intégralité de celui-ci).

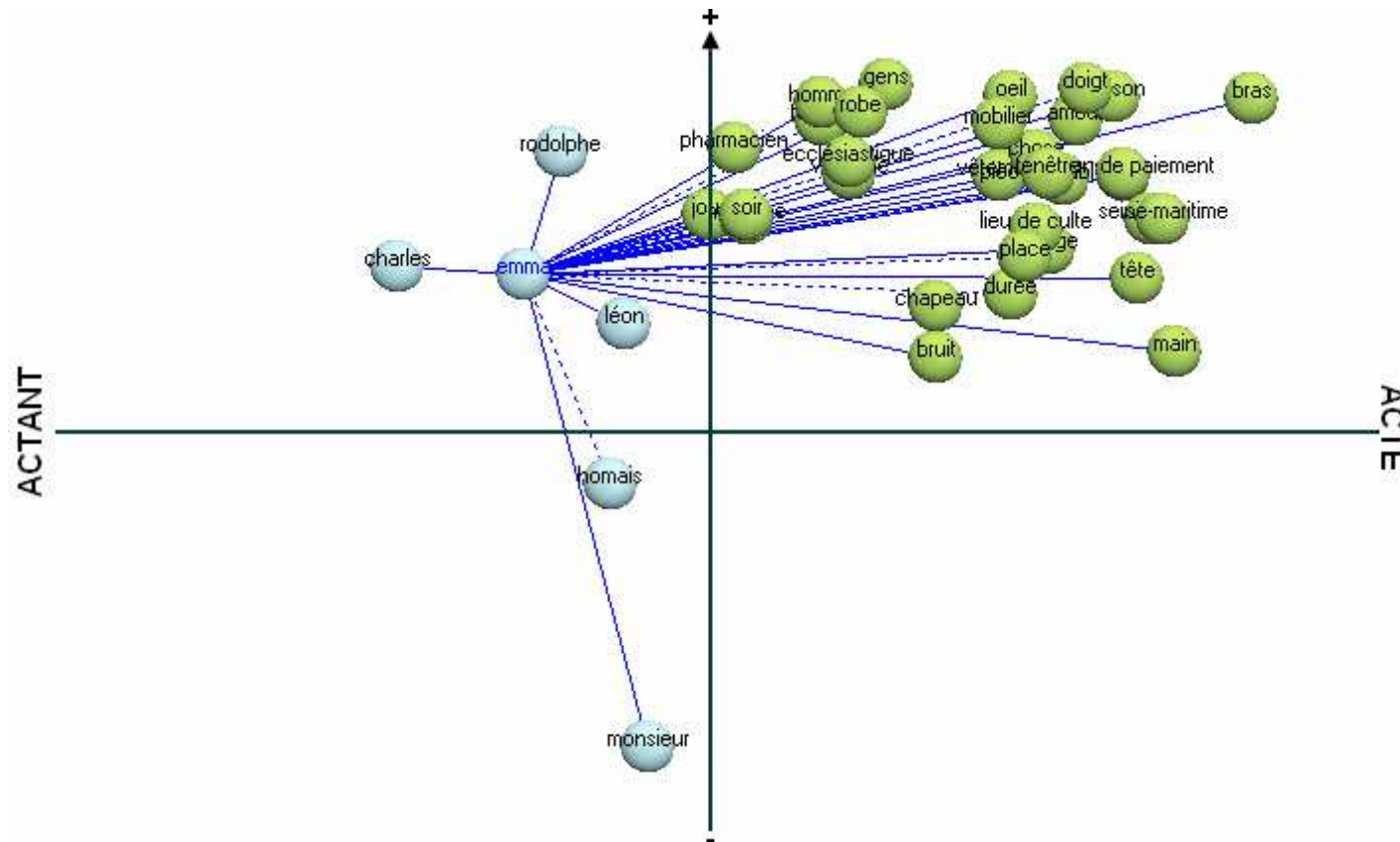
Un Episode correspond à une partie du texte dans lequel un certain nombre de Rafales se sont formées et terminées. Ce sont de grands blocs d'argumentation, représentatifs de la structure du discours observé.

Certaines catégories APD, ainsi que les pronoms personnels, sont affichés sur ces graphiques.

Sur cet exemple, on voit que le texte commence par "parler" d'élection, de problème et de citoyen, puis passe à un autre épisode évoquant le chômage, le travail et la sécurité sociale, etc.

Comment présenter certains résultats ? Des graphiques conçus comme des outils d'analyse (suite)

Le graphe des acteurs est une extension de l'APD, synthétisant fréquences d'occurrence, cooccurrences et acteurs :



Sur cet exemple on voit le résultat d'une analyse de *Madame Bovary* de Flaubert.

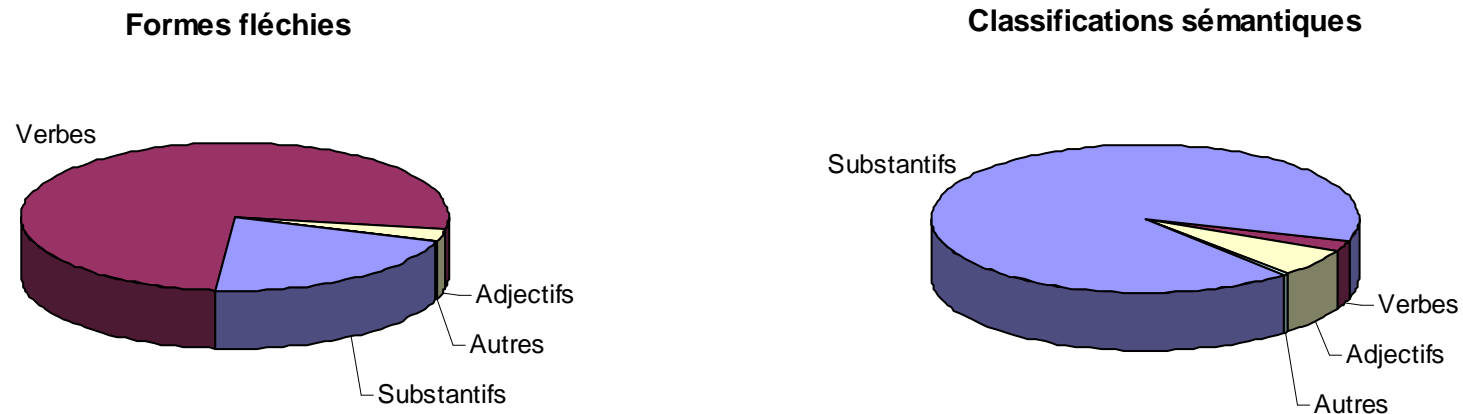
Tous les personnages principaux sont des actants et bénéficient d'une importante diversité de relation. On les voit en haut et à gauche du graphique.

Les autres référents, à droite, sont actés : on y retrouve des objets, des lieux ou des personnages secondaires.

L'axe vertical indique la « concentration de relations » pour chaque référence affichée. Il s'agit de pondérer la fréquence d'occurrence par le nombre de relations de cooccurrence différentes. Les traits indiquent les relations avec d'autres références.

Lexiques et réseaux sémantiques

Voici deux graphiques concernant la répartition des mots des dictionnaires, pour la langue française :

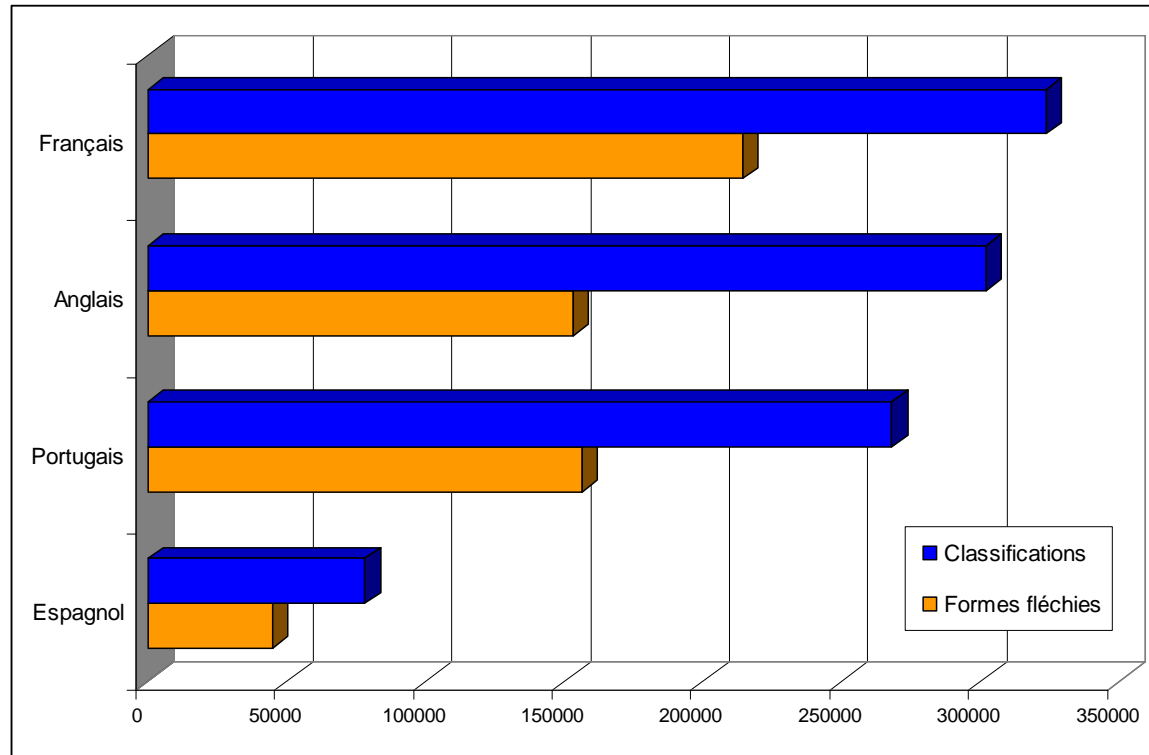


Le nombre de conjugaisons est élevé en Français, ce qui génère beaucoup de formes fléchies (800 000 théoriques). Toutefois les verbes sont peu nombreux (environ 8000 lemmes, sans compter en double les pronominaux), contrairement aux substantifs (environ 120 000 lemmes) qui font l'objet du plus grand nombre de classifications.

Les formes fléchies stockées dans le lexique de Tropes ont été extraites en analysant des corpus contenant des millions de documents (et des milliards d'occurrences de mots). Elles sont donc attestées. Les formes fléchies rares (par ex. plus-que-parfait du subjonctif) ou non-attestées sont gérées par un conjuguéur fléchisseur (qui corrige aussi certaines fautes d'orthographe).

Langues vivantes traitées

Des dictionnaires existent actuellement dans quatre groupes de langues, qui font fait l'objet de cinq versions différentes du logiciel : Anglais, Espagnol, Français et Portugais/Brésilien



Ce graphique indique, pour chaque langue, le nombre de formes fléchies attestées ainsi que le nombre total de classifications sémantiques (hors scénario).

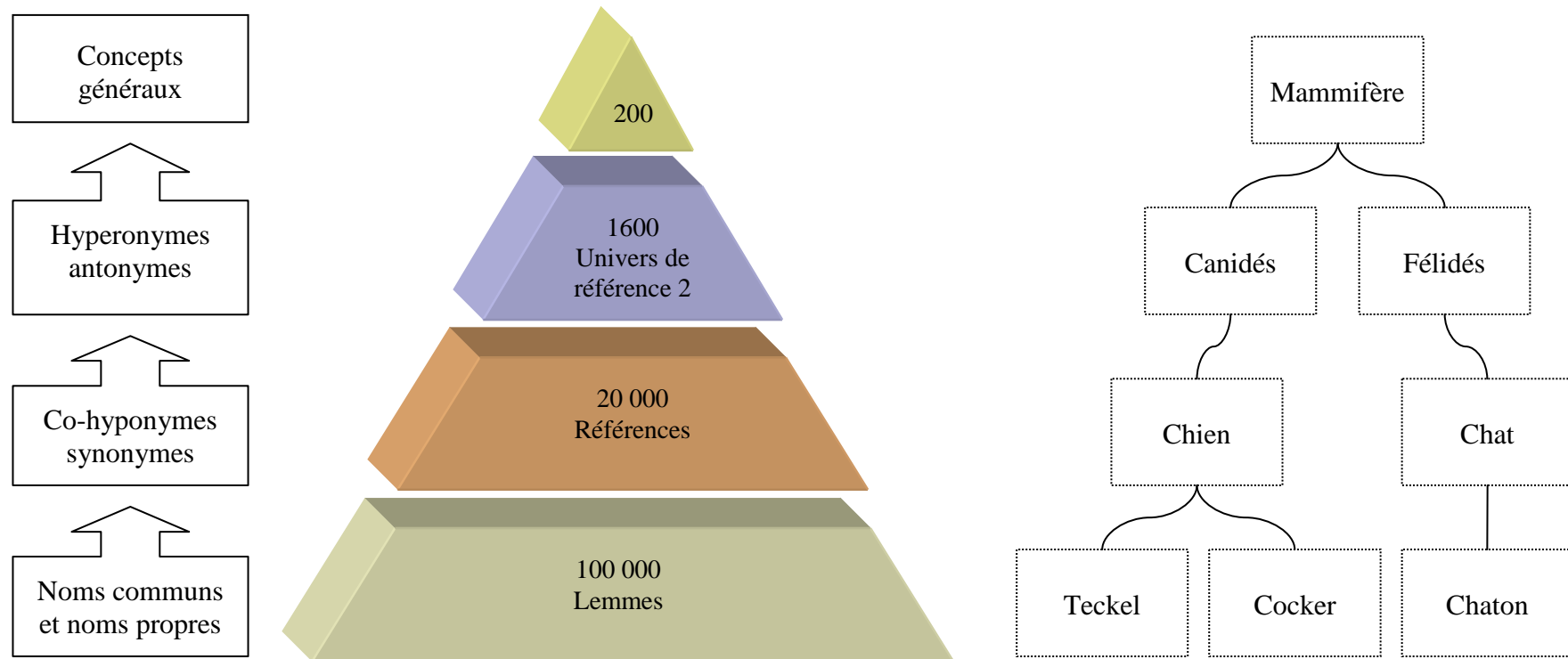
Il faut relativiser ces statistiques, sachant que ces langues ne sont pas toutes comparables. Par exemple, le nombre d'entrées canoniques anglaises dépasse les autres langues (qui ont beaucoup de formes fléchies, parce qu'elles ont des grammaires complexes).

Bien que la qualité des résultats ne soit pas la même, la version espagnole de Tropes gère en pratique (via le fléchisseur) autant de formes de verbes que la portugaise.

A titre de comparaison, la version espagnole actuelle de Tropes contient 4 fois plus de classifications et gère 30 fois plus de formes fléchies que le logiciel APD version 1992.

D'autres langues sont prévues (Roumain, Grec) ou existent à l'état de prototype (Allemand, Italien).

Analyse de la référence : des équivalents classés par hyperonymes



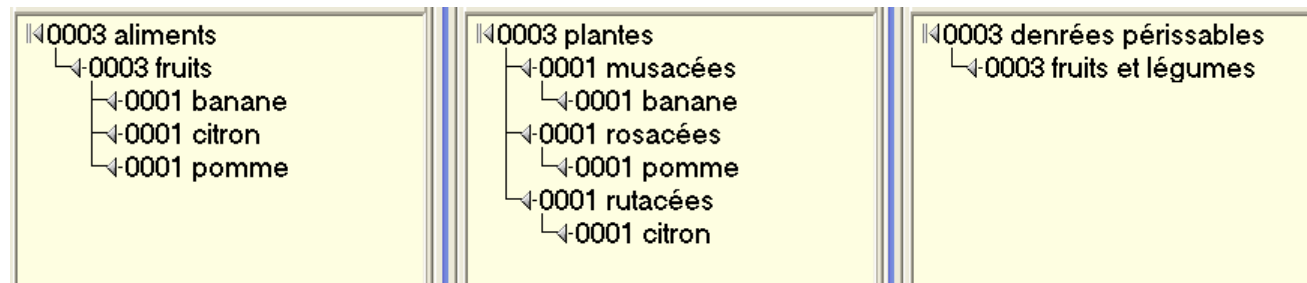
Le dictionnaire des équivalents de Tropes contient une triple classification des substantifs (Univers de référence 1 et 2, Références utilisées), ce qui permet une forte réduction du nombre de variables utilisées pour classer la référence. Ces classifications sont reprises dans l'arborescence des scénarios d'analyse. L'objectif est de prendre du recul, avant de passer à l'interprétation.

Le Scénario : un éditeur d'ontologies

Le Scénario est un outil interactif permettant de contextualiser l'analyse, qui répond à plusieurs objectifs :

Fonctionnalités	Objectifs
Compléter les classifications existantes et résoudre manuellement certaines ambiguïtés	Améliorer l'analyse ; Résoudre certains équivalents paradigmatiques
Proposer des constructions hiérarchisées, comme un thesaurus	Structurer le résultat
Permettre des classifications combinant substantifs, verbes et adjectifs	Elaborer d'autres modèles d'analyse
Définir plusieurs ontologies personnalisées en fonction des objectifs d'analyse	Test de plusieurs hypothèses

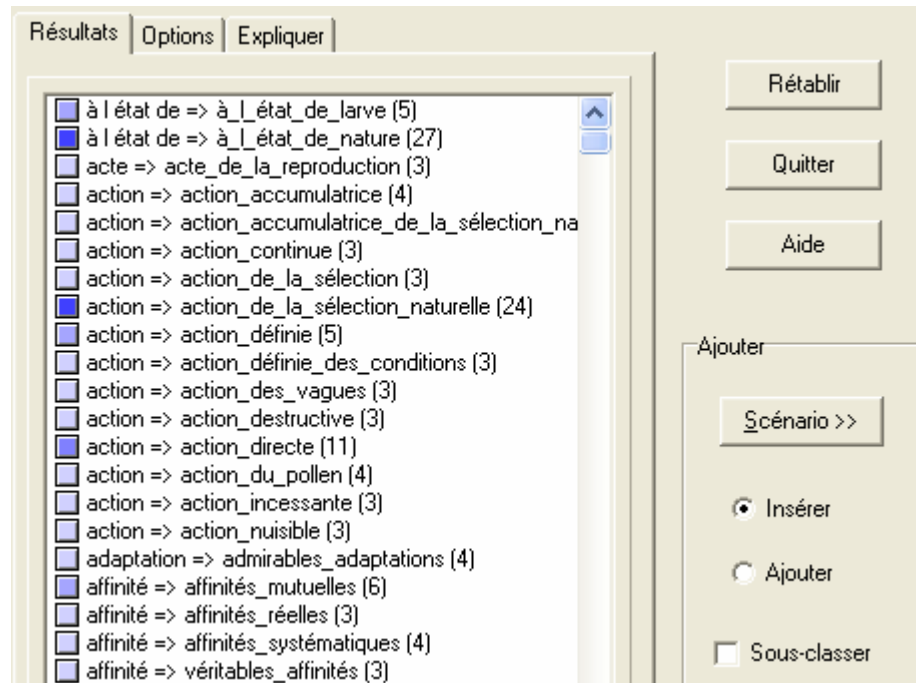
Par exemple, les mots « pomme », « citron » et « banane » peuvent, suivant le contexte, faire l'objet de trois classifications pertinentes, au sens courant, en botanique et suivant la nomenclature douanière.



Le Scénario n'impose pas les contraintes de l'APD. Par exemple, les substantif « syndicaliste », verbe « syndicaliser » et adjectif « syndical » peuvent être regroupés sous la référence [Syndicat], ce qui est correct du point de vue linguistique.

Une extraction terminologique couplée à l'analyse sémantique

Cet outil lexico-sémantique extrait du texte les mots composés ou expressions régulières (i.e. suite de termes répétés contenant au moins un substantif et cohérents d'un point de vue linguistique) qui peuvent présenter un intérêt pour l'analyse.

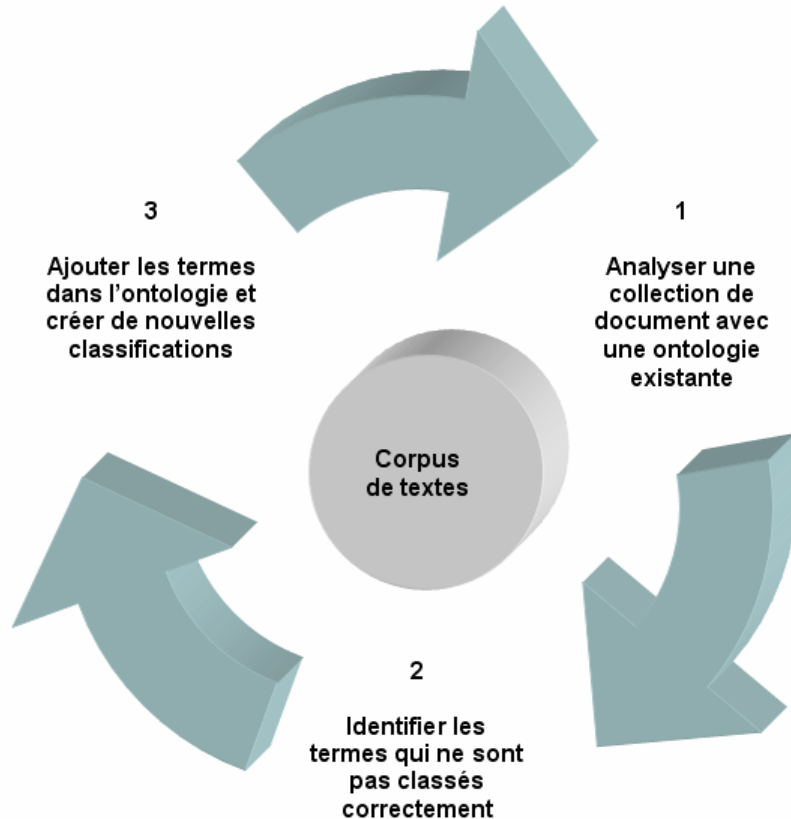


Sur cet exemple on voit une partie des résultats d'une extraction terminologique effectuée sur *L'origine des espèces* de Charles Darwin

Ce sont rarement de simples équivalents paradigmatiques.

L'extracteur terminologique permet, à la fois, d'enrichir rapidement les Scénarios du logiciel (en regroupant, par exemple, tous les sigles avec les expressions qui y correspondent) et d'obtenir une classification plus précise (en proposant, par exemple, de câbler les termes qui posent des problèmes d'ambiguïté et/ou qui peuvent « parasiter » l'analyse des cooccurrences (Relations)).

Une méthode récursive de construction d'ontologies



Si l'objectif est de disposer d'un plan de classement "exhaustif" pour un domaine, une approche pragmatique peut consister à extraire des textes de l'information sémantique (attestée) et à l'utiliser pour compléter une ontologie existante.

Cette méthode se fonde sur un processus récursif d'analyse, qui va partir d'une première classification (construite *a priori*) et boucler sur les trois étapes suivantes :

- 1 – analyser une collection de documents (corpus de test) représentative du sujet traité ;
- 2 – identifier tous les termes et expressions qui ne sont pas pris en compte dans la classification (et qui sont jugés pertinents par rapport à la problématique d'analyse) ;
- 3 – rajouter les termes pertinents dans la classification et repartir à l'étape 1 (autant de fois que nécessaire), en changeant éventuellement de corpus de test.

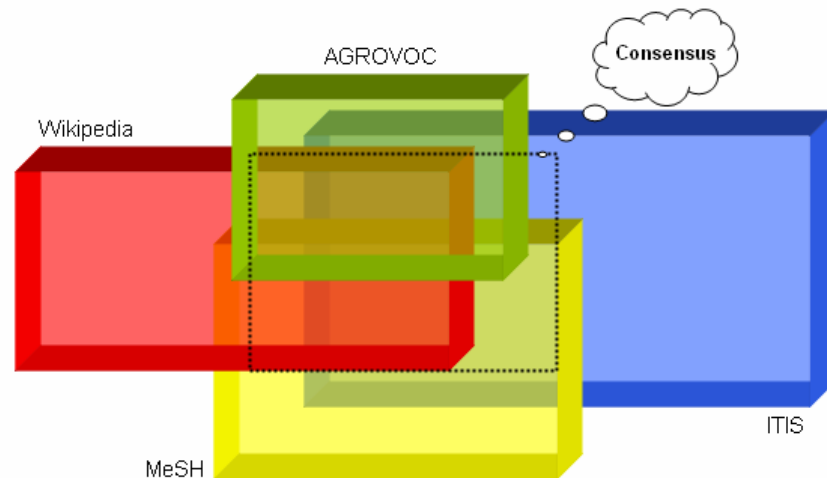
Sous certaines conditions, cette approche récursive peut être considérée comme terminée quand la classification reste stable lorsqu'on ajoute de nouveaux corpus.

Une méthode dérivée de ce qui précède peut servir à évaluer la qualité d'une ontologie ou un thesaurus, pour les faire évoluer.

Projet Agrovoc : un exemple d'évaluation de thesaurus scientifique

AGROVOC est un vocabulaire multilingue structuré (thesaurus) de la FAO (Food and Agriculture Organization) conçu pour couvrir la "terminologie de tous les domaines ayant trait à l'agriculture, à la pêche, à l'alimentation et aux domaines connexes".

Ce projet, réalisé en partenariat avec la FAO et le CIRAD, a dans un premier temps consisté à transformer le vocabulaire d'AGROVOC en réseau sémantique, puis à évaluer sa pertinence pour analyser des textes. Il a été ensuite décidé de compléter la classification par regroupement avec d'autres ontologies existantes. Ce qui a nécessité d'évaluer puis de fusionner plusieurs ontologies et d'arbitrer sur les parties qui pouvaient a priori être jugées comme les meilleures.



L'analyse des intersections entre les termes scientifiques communs à plusieurs ontologies (AGROVOC, ITIS, Mesh et Wikipedia) a donné trois ensembles distincts :

- 1 – les classifications communes à la majorité des ontologies (consensus) ;
- 2 – les classifications contradictoires (sans réel consensus) ;
- 3 – des classifications orphelines (qui n'existaient que dans une seule ontologie).

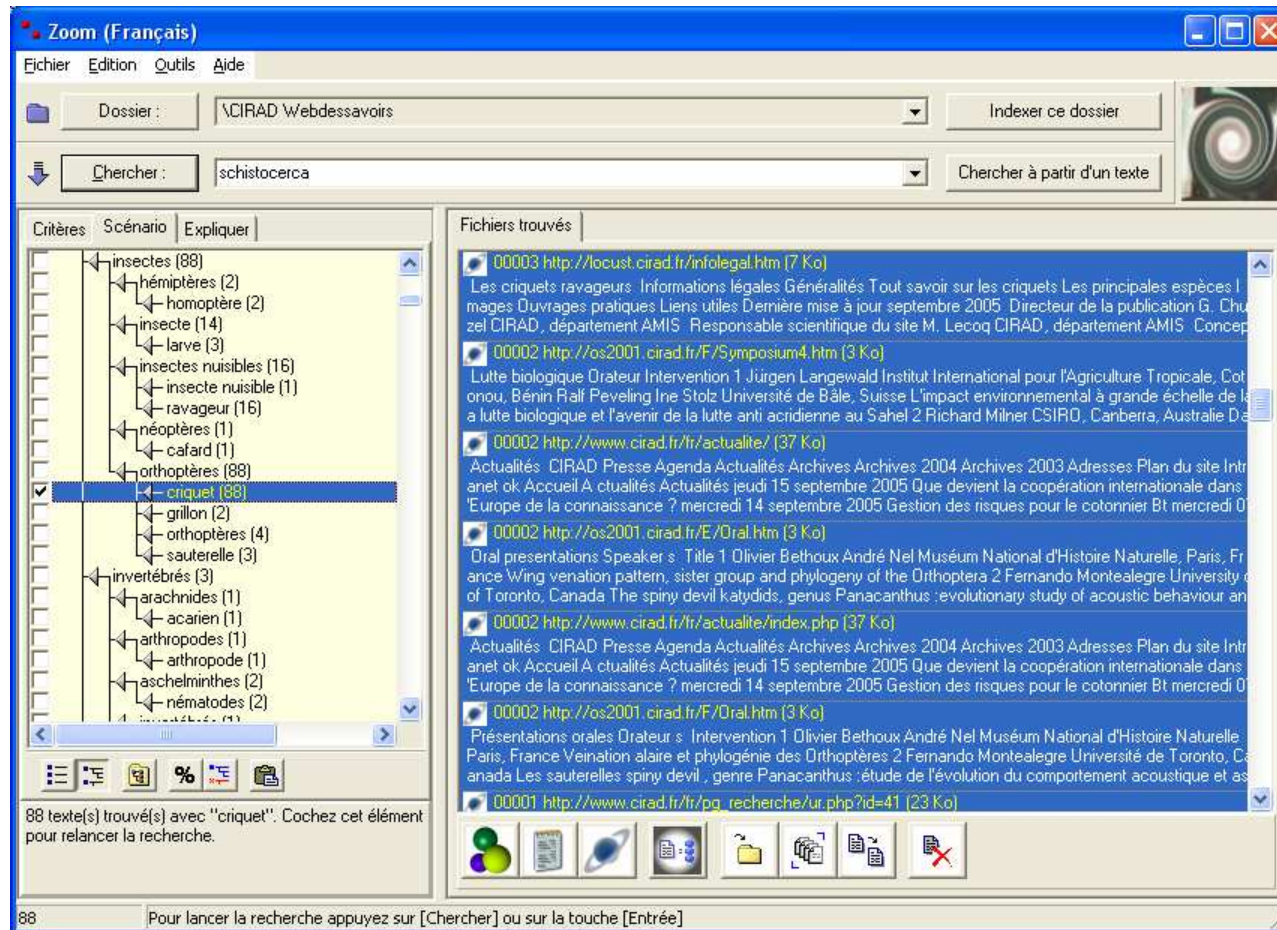
L'arbitrage a été rendu possible en constituant un corpus scientifique permettant de choisir entre telle ou telle classification.

ITIS (Integrated Taxonomic Information System) est une base de données multilingue d'informations taxonomiques concernant les plantes, les animaux, les champignons et les micro-organismes publiée par le US Department of Agriculture (USDA)

MeSH (Medical Subject Headings) est un thesaurus médical à vocabulaire contrôlé, de la National Library of Medicine (NIH).

L'encyclopédie Wikipedia contient des informations assez pertinentes dans le domaine des Sciences de la vie et ses taxinomies.

Réutilisation d'ontologies dans un moteur de classification : Zoom



Zoom utilise les scénarios de Tropes pour classer des structures de fichiers, des pages web ou des fonds documentaires

La capacité d'indexation de Zoom est conséquente : plusieurs millions de documents

Il est possible de regrouper les documents par références, puis de les analyser en bloc dans Tropes

Zoom permet donc de multiplier les analyses via une approche de génération dynamique de corpus

Dans l'exemple ci-dessus, une recherche sur le terme scientifique "schistocerca" a donné 88 documents (pages Web) inclus dans la référence "cricquet" (qui est un terme vernaculaire, hyperonyme de schistocerca).

Background scientifique et technique

Outils et fonctions d'analyse	Crédit
Catégories APD	Groupe de Recherche sur la Parole
Analyse propositionnelle	Groupe de Recherche sur la Parole
Style du texte	Patrick Charaudeau, Agnès Landré
Rafales	Mathieu Brugidou
Episodes	Pierre Molette
Propositions remarquables	Rodolphe Ghiglione, Pierre Molette
Analyse morphosyntaxique	Pierre Molette, Dan Caragea
Levée d'ambiguïté sémantique	Pierre Molette
Analyse de cooccurrence	Agnès Landré, Pierre Molette
Classification de la référence	John Lyons, Pierre Molette, Agnès Landré
Extraction terminologique et Scénario	Pierre Molette
Graphes	Pierre Molette
Dictionnaires et ontologies	Acetic, Cyberlex, Semantic-Knowledge

Références bibliographiques

- * ACETIC. Editeur de logiciels, et en particulier de Tropes et Zoom. <http://www.acetic.fr>
- * AGROVOC. FAO (Food and Agriculture Organization, Nations Unies). <http://www.fao.org/agrovoc/>
- * Brugidou, M. L'élection présidentielle : discours et enjeux politiques. Paris, L'harmattan, 1995.
- * Charaudeau P. Grammaire du sens et de l'expression. Paris, Hachette-Education, 1992.
- * CIRAD (Centre de coopération internationale en recherche agronomique pour le développement). Organisme scientifique spécialisé en agriculture des régions tropicales et subtropicales . <http://www.cirad.fr>
- * Ghiglione R., Landré A., Bromberg M., Molette P. L'analyse automatique des contenus. Paris, Dunod, 1998.
- * Ghiglione R., Kekenbosch C., Landré A. L'analyse cognitivo-discursive. Grenoble, Presses Universitaires de Grenoble, 1995.
- * Grevisse M., Goosse A. Le bon usage. Paris, Duculot, 1993.
- * ITIS (Integrated Taxonomic Information System). USDA (US Department of Agriculture). <http://www.itis.gov>
- * Le Quéau P., Brugidou M. La dynamique interne du récit. Paris, Cahier de recherche Crédoc numéro 124, 1998.
- * Lyons J. Sémantique linguistique. Paris, Larousse, 1980.
- * MeSH (Medical Subject Headings). National Library of Medicine (NIH, USA). <http://www.nlm.nih.gov/mesh>