# Le raisonnement analogique en lexicographie, son informatisation et son application au Réseau Lexical du Français

Sandrine Ollinger

UNIVERSITÉ DE LORRAINE

École doctorale Stanislas

# Le raisonnement analogique en lexicographie, son informatisation et son application au Réseau Lexical du Français

## THÈSE

présentée et soutenue publiquement le 15 décembre 2014

pour l'obtention du

**Doctorat de l'Université de Lorraine**

**Mention Sciences du langage**

par

**Sandrine Ollinger**

**Composition du jury**

| | | |
|---|---|---|
| **Sylvain Kahane,** | Professeur, Université Paris Ouest Nanterre | Rapporteur |
| **Mathieu Lafourcade,** | Maître de conférence, HDR, Université Montpellier 2 | Rapporteur |
| **Marie Candito,** | Maître de conférence, Université Paris Diderot | Examinatrice |
| **Bruno Gaume,** | Chargé de recherche, CLLE-ERSS, Toulouse | Examinateur |
| **Jean-Marie Pierrel,** | Professeur, Université de Lorraine | Examinateur |
| **Alain Polguère,** | Professeur, Université de Lorraine | Directeur |

Analyse et Traitement Informatique de la Langue Français — UMR 7118

atilf ANALYSE ET TRAITEMENT INFORMATIQUE DE LA LANGUE FRANÇAISE cnrs UNIVERSITÉ DE LORRAINE

*À Anne Joséphine.*

# Remerciements

Merci, Alain, d'avoir accepté de diriger ce travail de thèse, de m'avoir proposé un sujet passionnant et d'avoir encadré mes recherches avec patience et discernement. Merci pour tes connaissances, tes convictions et l'énergie débordante que tu consacres à la réalisation du RL-fr.

Merci à Mathieu Lafourcade et Sylvain Kahane d'avoir accepté d'être rapporteurs de ce travail de thèse. Merci à Marie Candito d'avoir accepté d'en être examinatrice. Merci à Bruno Gaume et Jean-Marie Pierrel d'avoir accepté d'en être examinateurs.

Merci à François Yvon d'avoir accepté de me rencontrer au printemps 2013 pour discuter de la formalisation de l'analogie et pour ses précieux conseils.

Merci, Bruno, pour la conversation passionnante que nous avons eue à l'automne 2013 sur les graphes lexicaux et pour l'intérêt que tu as alors porté à mes travaux. Merci également aux autres membres de la Proxteam toulousaine, à Emmanuel Navarro pour son script d'analyse topologique et ses clefs d'interprétation, à Yann Desalle pour nos échanges scientifiques, qui, je l'espère, s'amplifieront à l'avenir et à l'ensemble de l'équipe pour leurs travaux captivants.

Merci, Jean-Marie, de m'avoir encouragée à entreprendre ce travail et de m'avoir aidée à concilier mes activités d'ingénieur d'études et de doctorante.

Merci à l'ensemble de l'équipe de lexicographes du projet RELIEF de m'avoir accueillie et d'avoir toujours été d'une grande disponibilité pour répondre à mes questions et m'apporter les précisions nécessaires à la bonne exploitation du RL-fr. Merci particulièrement à Anaïs, Candice, Christelle, Delphine et Sandrine, celles que j'ai le plus sollicitées et qui m'ont chaleureusement intégrée à tous ces petits à côté qui ont participé à faire de cette aventure une expérience humaine des plus enrichissantes. Je vous souhaite de connaître un franc succès dans la création d'$\mathbf{Oper_1}$ et de continuer longtemps à mettre vos talents au service de la lexicographie et de la terminologie.

Merci, Mi Hyun, pour nos déjeuners de doctorantes débordées et pour le cahier que tu m'as offert au moment opportun pour accompagner mes premières expériences de regroupement de sous-graphes analogues.

Merci à Nabil Gader pour le minutieux travail d'implémentation de l'éditeur lexicographique Dicet qu'il réalise et pour la disponibilité dont il a toujours fait preuve à mon égard.

Merci à tous mes collègues de l'ATILF, présents et passés. Travailler avec vous depuis bientôt huit ans est une chance exceptionnelle. Merci pour vos activités scientifiques, pour votre bonne humeur, pour vos coups de gueule, pour votre créativité, pour votre curiosité, pour votre solidarité et pour le maintien de votre esprit de service public en ces temps difficiles. Merci pour les moments d'échanges passés autour d'un café, d'un gâteau, d'un barbecue, tout autant qu'autour de tables de réunion. Merci pour le partage de vos connaissances, pour votre disponibilité, pour votre ouverture d'esprit et votre soif de collaborations. Merci pour vos encouragements, votre patience et votre soutien sans faille. Un merci spécial pour toi, Étienne, qui ne désespères pas de faire de moi une *vraie* informaticienne, parfaitement autonome. Un autre pour toi, Mathieu, avec qui j'ai réalisé mes premières collaborations scientifiques autour de la néologie. Merci de m'avoir incitée à rédiger avec toi mes premiers articles et à présenter les résultats de nos travaux communs. Merci à Évelyne et à toi d'avoir toujours accordé de l'intérêt à ce que la petite ingénieure en informatique avait à dire sur les expériences que vous réalisiez et les analyses qui en découlaient. Merci encore, Évelyne, pour les cours que nous avons montés et assurés ensemble, ça y est, on peut recommencer ! Merci, Cyril, d'être le collègue de bureau idéal. Merci, Iveta, Coralie et Franziska, qui avez fait appel à moi pendant vos doctorats respectifs. Vous aussi, vous avez joué un rôle important pour que je décide d'entamer le mien.

Merci à mes amis et à ma famille, que j'ai quelque peu délaissés ces derniers mois, mais qui ont su se montrer persévérants. Un merci particulier aux *meufs* Marie, Flora, Zou et Lucile pour les séances de gommes des premiers mois de thèse et pour tout le reste. Un autre à Émilie, Julien et Pierre pour leurs invitations multiples et variées, malgré mes nombreux refus. Encore un aux amies de longue date Amandine, Aurélie, Manue et Muriel, ainsi qu'à Sylvain. Un autre à mon grand frère, jamais très loin de l'autre côté du tchat. Un dernier à Christophe et à Noam. Merci à tous ceux que je n'ai pas mentionnés, mais qui ont participé, d'une façon ou d'une autre, à ce que je débute ce doctorat et à ce que je parvienne à le terminer.

Merci.

# Résumé

La lexicographie contemporaine, en mettant à profit les avancées théoriques et pratiques de l'informatique et de la linguistique, s'est affranchie de l'organisation linéaire imposée par les ouvrages papier. Elle s'est attachée à définir de nouveaux modèles de description et met aujourd'hui à disposition de la communauté des ressources formelles offrant de multiples possibilités d'exploitations automatiques. Cette thèse concentre son attention sur le modèle des systèmes lexicaux proposé par la Lexicographie Explicative et Combinatoire. Plus précisément, elle s'intéresse au Réseau Lexical du Français, en cours de développement. En tant que système lexical, cette ressource est un graphe lexical monolingue. Elle est constituée d'un ensemble de sommets, les unités lexicales du français, entre lesquels sont encodées de nombreuses relations, en grande majorité syntaxico-sémantiques.

La présente thèse pose les bases d'une exploration de cette ressource lexicographique par raisonnement analogique. Elle débute par une revue sélective de la formalisation et de l'informatisation de l'analogie en traitement automatique des langues, dans le cas précis de l'étude du lexique. Elle définit ainsi le principe de l'exploration réalisée comme un regroupement de structures unifiables. Les sommets du graphe lexical s'apparentent alors à des objets disposant d'un certain nombre d'Attributs, disponibles dans les description lexicographique qu'ils encapsulent. Ils entretiennent des Relations, représentées par les arcs.

Une réflexion est menée sur la nature des différents éléments composant le réseau et sur les différents rapports qu'ils entretiennent. Elle est réalisée en prenant en compte l'évolution de la ressource sur une période de trente mois. Elle est accompagnée d'une analyse topologique, qui met en avant des propriétés proches de celles des graphes petit monde.

Deux séries d'expériences exploratoires sont ensuite réalisées. La première d'entre elles permet de conforter l'idée selon laquelle la formalisation en œuvre dans la ressource permet de détecter automatiquement des analogies conformes à l'intuition des locuteurs. Elle met en avant la possibilité de réaliser différents types d'explorations par raisonnement analogique, en fonction des points d'entrée choisis et des éléments d'informations comparés. Elle montre également l'apport de telles explorations en terme de vérification de la cohérence du réseau et d'émergence de règles lexicales. La seconde série d'expériences se concentre autour de la notion de configurations de dérivations lexicales. Elle montre comment le regroupement de sous-graphes analogues met en avant l'existence de connexions lexicales récurrentes à travers la ressource.

L'état d'avancement de la ressource exploitée ne permet pas d'obtenir des règles

et des modèles aboutis. Les résultats obtenus sont toutefois encourageants. Les observations réalisées nous amènent à considérer l'analogie comme un guide permettant de s'assurer de la bonne qualité de la représentation du lexique proposée par une ressource. Elle permet également d'acquérir automatiquement des connaissances sur son organisation. De telles connaissances permettent d'identifier des phénomènes linguistiques et d'instrumenter l'activité lexicographique.

# Table des matières

# Table des figures

# Liste des tableaux

# Liste des abréviations, symboles et conventions d'écriture

## Liste des abréviations

| | |
|---|---|
| CFC | Composante Faiblement Connexe |
| CG | Caractéristique Grammaticale |
| DULB | Description d'une unité lexicale de base |
| DULS | Description d'une unité lexicale secondaire |
| FL | Fonction Lexicale |
| FP | Forme Propositionnelle |
| méta-pdd | méta-partie du discours |
| RL-fr | Réseau Lexical du Français |
| $sim_a$ | Similarité d'Attributs |
| $sim_r$ | Similarité de Relations |

## Symboles et conventions d'écriture

| | |
|---|---|
| VOCABLE$_{\text{indice}}^{\text{exposant}}$ : | SUJET$_{\text{N}}^{1}$ |
| LEXÈME$_{\text{indice}}^{\text{exposant}}$ **numéro** : | SUJET$_{\text{N}}^{1}$ **III** |
| [*illustration de sens*] : | [*Il collectione des sujets en résine.*] |
| ⌜ LOCUTION ⌝ : | ⌜ PLANCHER DES VACHES ⌝ |
| **fonction lexicale** : | **Magn** |
| `caractéristique grammaticale` : | `nom commun` |
| `table flexionnelle prototypique` : | `chat` |
| étiquette sémantique : | ensemble de chose physique |
| CLASSE_SÉMANTIQUE : | CHOSE_[PHYSIQUE] |
| `forme propositionnelle` : | ∼ `qui sert à X pour Y` |
| proportion analogique : | $A : B :: C : D$ |

# Introduction générale

## Contexte

La lexicographie contemporaine, telle que la décrivent Atkins (1996), Selva et al. (2003) et Spohr (2012), a su mettre à profit les avancées théoriques et pratiques de l'informatique et de la linguistique. Elle s'est affranchie de l'organisation linéaire imposée par les ouvrages papier, pour s'intéresser à la conception de bases de données lexicales complexes. Elle s'est ainsi attachée à définir de nouveaux modèles de description et met aujourd'hui à disposition de la communauté des ressources formelles offrant de multiples possibilités d'exploitations automatiques.

Cette thèse se concentre sur le modèle des systèmes lexicaux proposé par la Lexicographie Explicative et Combinatoire (Polguère, 2009, 2014a,b). Plus précisément, elle s'intéresse au Réseau Lexical du Français. En tant que système lexical, cette ressource est un graphe lexical monolingue. Elle est constituée d'un ensemble de sommets, les unités lexicales du français, entre lesquels sont encodées de nombreuses relations, en grande majorité syntaxico-sémantiques.

Le Réseau Lexical du Français est actuellement en cours de développement au laboratoire d'Analyse et Traitement de la Langue Française (ATILF), où a été réalisé le présent travail de recherche, depuis juin 2011. Il bénéficie du soutien de l'Agence de Mobilisation Économique de la Région Lorraine et du Fond Européen de Développement Régional Lorrain dans le cadre du projet RELIEF, *REssource Lexical Informatisée d'Envergure sur le Français*[1]. Dans ce contexte, une équipe d'une dizaine de lexicographes travaillent à la description du lexique du français. Ce travail s'organise en tâches lexicographiques et est accompagné de l'élaboration d'un éditeur lexicographique, en partenariat avec la société privée MVS de Saint-Dié[2].

Comme nous le verrons par la suite, la première de ces tâches lexicographiques a été la mise au point d'une nomenclature de base, selon les principes décrits dans Polguère et Sikora (2013). Les autres tâches ne s'organisent pas nécessairement de manière consécutive. Elles peuvent occuper l'intégralité de l'équipe, comme dans le cas de l'encodage de l'ensemble des dérivés sémantiques proches de la nomenclature initiale, ou une sous-équipe de lexicographes, comme la gestion de l'ontologie d'étiquettes sémantiques ou l'encodage des informations morphologiques flexionnelles. La plupart d'entre elles sont accompagnées de la conception et de l'implémentation de nouvelles fonctionalités dans l'éditeur lexicographique. Le travail de description

[1] `http ://www.atilf.fr/spip.php?article908`
[2] `http ://www.mvs.fr`

lexicographique ne s'organise donc pas de façon linéaire, entrée après entrée, mais d'une façon plus organique, facette après facette.

La présente thèse s'intéresse à la cohérence de cette ressource. Elle part du principe qu'en s'appuyant sur les aspects formels des nouveaux modèles de description lexicographique, il est possible, d'une part, de vérifier la cohérence des ressources développées et, d'autre part, d'acquérir automatiquement des connaissances sur l'organisation du lexique. De telles connaissances permettraient alors d'identifier des phénomènes linguistiques, qu'il serait possible d'exploiter pour instrumenter[3] l'activité lexicographique. D'autres applications seraient également envisageables, dans le cadre d'une instrumentation plus générale de la recherche en lexicologie et d'activités pédagogiques autour du lexique.

## Exploration par raisonnement analogique

Nous avons eu l'occasion, au cours de ce travail de recherche, d'assister à de nombreuses réunions de travail de l'équipe de lexicographes et de suivre avec attention l'organisation de leur activité. Nous avons alors pu constater l'importance des processus de comparaison dans leur manière de fonctionner. Que ce soit pour déterminer si un signifiant linguistique correspond à une ou plusieurs unités lexicales, choisir entre différentes étiquettes sémantiques ou encore décider de la création d'un nouveau type de relation à encoder, les lexicographes procèdent régulièrement par l'analyse conjointe de différentes situations.

Le raisonnement analogique qu'ils mettent ainsi à l'œuvre a fait l'objet de nombreux travaux de formalisation et d'informatisation aux cours des trente dernières années. Ainsi, dans le cadre de la psychologie cognitive, les travaux de Gentner (1983) ont ouvert la voie à de nombreuses études sur l'analogie en tant que procédé de compréhension d'une nouvelle situation par exploitation des connaissances acquises dans une situation antérieure. Cet aspect intéresse également la recherche en intelligence artificielle, qui a d'ores et déjà développé de nombreuses méthodes de résolution de problèmes nouveaux à partir de l'analyse de cas similaires connus (Aamodt et Plaza, 1994). En traitement automatique des langues, des méthodes et des outils spécifiques ont été mis au point, adaptés à l'analogie entre chaînes de symboles. Ils permettent la réalisation de tâches variées, telles que la génération automatique de phrases (Lepage, 2003), le changement de niveau de représentation d'analyses linguistiques (Stroppa, 2005) ou la translitération de noms propres de langue anglaise vers le chinois (Langlais et Yvon, 2014). L'analogie sémantique entre unités lexicales a également été abordée (Turney, 2006 ; Veale, 2004).

Le présent travail de thèse propose une adaptation de telles méthodes à la tâche de délimitation de sous-ensembles cohérents dans un modèle lexicographique donné. En nous appuyant sur la formalisation du lexique inhérente au modèle des systèmes lexicaux, nous développons l'idée qu'une exploration de ressources lexicales par raisonnement analogique permet d'en vérifier la cohérence et constitue une étape importante vers l'abstraction de connaissances nouvelles.

[3]Nous parlons ici d'instrumentation au sens de Habert (2005).

## Organisation de la thèse

Cette thèse s'organise en cinq chapitres, une introduction générale et une conclusion. Elle est accompagnée d'une annexe présentant une partie des fonctions lexicales Sens-Texte, afin de faciliter la lecture des analyses réalisées.

Le premier chapitre offre une revue sélective de la formalisation et de l'informatisation de l'analogie en traitement automatique des langues, dans le cas précis de l'étude du lexique. Il définit ainsi le principe d'exploration par raisonnement analogique comme un regroupement de structures unifiables. Les sommets du graphe lexical s'apparentent alors à des objets disposant d'un certain nombre d'Attributs, disponibles dans leur description lexicographique. Ils entretiennent des Relations, représentées par les arcs.

Le second chapitre consiste en une réflexion sur la nature des différents éléments composant le Réseau Lexical du Français et sur les différents rapports qu'ils entretiennent. Il prend en compte l'évolution de la ressource sur une période de trente mois. Il s'achève sur la présentation d'une analyse topologique du réseau, qui met en avant la proximité de ses propriétés avec celles des graphes petit monde.

Le troisième chapitre présente une première série d'expériences exploratoires, basée sur la comparaison de descriptions lexicographiques deux à deux. Il permet de confirmer l'hyptohèse selon laquelle la formalisation en œuvre dans la ressource permet de détecter automatiquement des analogies conformes à l'intuition des lexicographes. Il met également en avant la possibilité de réaliser différents types d'explorations par raisonnement analogique, en fonction des points d'entrée et des éléments d'information comparés. Enfin, il montre l'apport de telles explorations en termes de vérification de la cohérence du réseau et d'émergence de règles lexicales.

Le quatrième chapitre propose d'abandonner la comparaison d'unités lexicales deux à deux pour s'intéresser à des petits ensembles cohérents réguliers. Dans ce cadre, nous définissons la notion de configurations de dérivations lexicales, en tant qu'abstraction de structures de dérivations syntaxiques et sémantiques récurrentes. Le chapitre présente une première expérience de regroupement de sous-graphes par raisonnement analogique, réalisée à partir de l'ensemble des sous-parties du Réseau Lexical du Français constituant des graphes indépendants révélés lors de son analyse topologique.

Le cinquième chapitre décrit l'adaptation de la méthode mise au point au cours du quatrième chapitre à l'exploration de sous-graphes immergés dans la structure du Réseau Lexical du Français. Il détaille une dernière série d'expériences permettant de valider la pertinence de notre démarche. Il se termine par l'amorce d'une réflexion sur l'automatisation d'une abstraction de configurations de dérivations lexicales.

Enfin, la conclusion générale revient sur les principaux constats effectués au cours de ce travail de recherche et présente un ensemble de perspectives qui s'esquissent à présent.

# Chapitre 1

# Raisonnement analogique

## Sommaire

# Introduction

> L'analogie vulgaire c'est la simple similitude de la perception : quelque chose est analogue à quelque chose d'autre. Si vous voulez c'est la similitude de la perception ou l'analogie de l'imagination, en gros ça se tient.
>
> Deleuze (1974)

Deleuze, qui s'intéresse à la métaphysique et, plus précisément, à la nature de l'être, introduit ainsi la notion d'*analogie vulgaire*, qu'il oppose à une analogie scientifique ou technique. Nous préférons, pour notre part, parler d'une *analogie de sens commun*, qui s'oppose à une *analogie formalisée*. Dans un cadre lexical, dire « *aboyer* est analogue à *beugler* », ça peut être une simple analogie de sens commun. Nous percevons une similitude entre ces mots. Nous savons, en tant que locuteur du français, que chacun d'entre eux a, a minima, un sens correspondant à l'action de pousser un cri pour un certain animal (*Son chien aboie.* ∼ *Cette vache beugle depuis l'aube.*) et un sens désignant une façon particulière de s'exprimer (*Il aboie des ordres depuis une heure.* ∼ *Elle ne chante pas, elle beugle.*). Lorsque nous consultons un dictionnaire, nous nous attendons naturellement à ce que cette similitude perçue soit confirmée ou infirmée ; quoi qu'il en soit, qu'elle soit précisée par les descriptions lexicographiques mises à notre disposition.

Ce premier chapitre établit les bases de notre réflexion sur l'analogie lexicale. Après une brève illustration de l'analogie de sens commun, la notion d'analogie formalisée y est introduite, dans un premier temps d'un point de vue mathématique, en tant que proportion ; dans un second temps, d'un point de vue psychologique, en tant que mécanisme de raisonnement. La question de l'analogie en œuvre dans l'organisation du lexique est ensuite discutée, par l'intermédiaire d'un panorama sélectif de son exploitation en traitement automatique des langues.

## 1.1 Analogie de sens commun

Ce que nous entendons par analogie de sens commun, c'est l'expression d'une certaine ressemblance. *Quelque chose est analogue à quelque chose d'autre* et si un locuteur l'exprime ainsi, c'est qu'il part du principe que son interlocuteur connait suffisemment ce *quelque chose d'autre* pour en tirer des conclusions sur le *quelque chose*. Les trois exemples suivants, tirés du corpus FrWac (Baroni et al., 2009), illustrent cet usage de l'adjectif ANALOGUE :

> Pour que les élèves n'aient pas trop d'outils à manipuler nous voulions avoir un serveur e-mail **analogue** à ce que l'on trouve sur Internet.
>
> `http ://irh.unice.fr/spip.php?article301`

> Ce que j'aime chez eux c'est que les vêtements sont tous très tendances, parce qu'être écolo ne veut pas dire ressembler à un hippie attardé. Les prix sont tout à fait **analogues** à des vêtements de moyenne gamme (25 euros pour leur top cintré très mode , par exemple) et en plus d' être

bios, ces vêtements sont issus de commerce équitable, ce qui ne gâche rien.

`http ://www.ecoblog.fr/Consommation-alternative/2006/07`

C'est **analogue** à l'utilisation de la balise link en HTML pour importer une CSS et la syntaxe est semblable : ?xml version=”1.0” ? ?xml-stylesheet href=”rss.css” type =”text/css” ? À la différence du HTML, aucune supposition n'est faite sur le formatage des éléments XML de la part du processeur.

`http ://mozilla.tlk.fr/doc06.php`

L'analogie de sens commun recouvre ainsi à la fois un processus de comparaison et son résultat. Les exemples suivants, extraits d'un corpus d'archives du mensuel *Le Monde Diplomatique* de 1994 à 1998, illustrent ces deux aspects. Nous pouvons y voir qu'il est parfois difficile de déterminer avec précision si l'occurrence du nom ANALOGIE qui s'y trouve dénote le processus de comparaison ou son résultat.

C'était, bien entendu, une généralisation, puisque « fascisme » ne désigne, au sens propre, que le système de l'Italie mussolinienne entre 1922 à 1943. Mais un certain nombre d'**analogies** ont conduit à mettre sous cette rubrique le régime de l'Allemagne nazie d'un côté, les régimes de conservatisme autoritaire d'Antonio Salazar, du général Franco ou du maréchal Pétain de l'autre.

Le Monde Diplomatique 1994

Sous le III$^{e}$ Reich, il existe une masse de tableaux qui sont tout simplement représentatifs de la traditionnelle peinture de genre. L'idéologie nazie disparue, ils restent comme des croûtes, et rien d'autre. Ce sont ces tableaux-là qui offrent, dans leur forme, des **analogies** avec la peinture soviétique des années 30.

Le Monde Diplomatique 1994

« On est devenus les Arabes des Luxembourgeois », déplore Mme Brandstaedt, ancienne femme de ménage dans une banque. En Lorraine, ce genre d'**analogie** ne laisse pas insensible. Après la fin de la seconde guerre mondiale, la cohabitation entre une population elle-même issue de vagues successives d'immigration et les derniers arrivants se passait sans encombres.

Le Monde Diplomatique 1997

Enfin, dans certains emplois, l'analogie est considérée comme un processus de création. Le dernier exemple, ci-dessous, tiré du *corpus Chambers-Rostand du français journalistique*[1], illustre cet usage.

[1] `http ://ota.ahds.ac.uk/desc/2491`

> Josse de Momper, peignant ses allégories des saisons, fait donc surgir des têtes de vieillards barbus de la roche, avec une tour en guise de nez et des broussailles pour chevelure. Max Ernst, trois cents ans plus tard, découvre un sphinx dans l'empreinte d'une éponge, un oiseau dans le corail, un dragon dans la mousse. Ainsi fonctionnent perception et création : par associations, par **analogies**, par glissements.
>
> 1_M_C_210403 \LeMonde\2003

Nous ne nous attarderons pas davantage sur cette analogie de sens commun. Si cette analogie entretient sans aucun doute un lien sémantique étroit avec la notion d'analogie qui intéresse les mathématiciens, les philosophes et les psychologues, nous ne nous engageons pas sur le chemin d'une telle analyse. Nous concentrons plutôt, dès à présent, notre attention sur ce que nous considérons comme une *analogie formalisée*, à travers les notions centrales de *proportions* et de *raisonnement* analogiques.

## 1.2 Proportions analogiques

La proportion analogique se note traditionnellement $A : B :: C : D$[2] et s'énonce « A est à B ce que C est à D ». Elle trouve son origine dans les travaux d'Euclide (III$^{e}$ s. av. J.-C.)[3]. Elle a, par la suite, été exploitée directement en mathématiques et de manière plus ou moins directe dans d'autres disciplines, telle que la philosophie, la psychologie, les sciences de l'éducation et la linguistique. La présente section rappelle l'origine de la notion d'analogie et son application en mathématiques élémentaires, avant d'en résumer la proposition de formalisation réalisée par Lepage (2003).

### 1.2.1 Proportion

Euclide introduit la notion d'*analogie* dans le cadre d'une étude des éléments géométriques, et ce, en deux temps. Dans un premier temps, il définit — troisième définition de l'élément cinquième — la notion de *raison*, en tant qu'« habitude de deux grandeurs de même genre, comparées l'une à l'autre selon la quantité ». Il détaille alors la distinction entre quantité et qualité et insiste sur la *similarité de genre* entre les quantités comparées. Ainsi, il s'agit du rapport entre deux valeurs comptables directement comparables (deux nombres, deux superficies, deux longueurs de ligne, etc.). Différents types de raisons sont ensuite discutés, dont nous ne conservons ici, pour les besoins de l'exemple, que la *raison multiple*, correspondant aux cas où la première valeur de la comparaison contient plusieurs fois la seconde :

> Soit deux lignes A et B, A mesurant 6 centimètres, B 3 centimètres, la raison de A à B est dite raison multiple, A contenant deux fois B.

[2]Différentes notations se rencontrent dans la littérature, telles que $A : B = C : D$ ou $A : B \doteq C : D$, que nous ne discuterons pas dans le cadre de ce travail.

[3]Nous parlons ici de la version traduite par Henrion (1632) des *Quinze livres des éléments géométriques*, consultée sur le site de Gallica, le 11 juillet 2014. `http ://gallica.bnf.fr/ark :/12148/bpt6k68013g`

Dans un second temps, il définit — quatrième définition de l'élément cinquième — la notion de *proportion*, en tant que *similitude de raisons*[4].

> Tout ainsi que la comparaison de deux quantités entre elles est dite raison, ainsi la comparaison et ressemblance de deux ou plusieurs raisons entre elles, est dite proportion : comme si la raison de A à B, est semblable à la raison de C à D, l'habitude d'entre ces raisons sera dite proportion. Et c'est ce que les Grecs appellent **analogie**, et quelques Latins **proportionnalité**[...].
>
> Euclide (III$^{e}$ s. av. J.-C.)

À partir de cette définition, reprenons les lignes A et B. Si nous comparons leur raison à celle des lignes C et D, mesurant respectivement 12 et 24 centimètres, nous observons que A et C contiennent deux fois, l'une B, l'autre D. Nous pouvons alors énoncer l'analogie suivante, où la raison équivaut au double :

> La raison de A à B est semblable à la raison de C à D.

La lecture de la suite de l'élément cinquième nous apprend que deux autres analogies sont directement énonçables à partir de celle-ci. L'une, par *permutation* — douzième définition de l'élément cinquième — « comme A est à B, ainsi C est à D : Donc en permutant comme A sera à C, ainsi B à D. Et est à noter qu'en cette manière d'argumenter, les quatre grandeurs doivent être de même genre. ». Nous observons alors que la longueur de la ligne A est à la longueur de la ligne C ce que la longueur de la ligne B est à la longueur de la ligne D : son quart[5]. L'autre, par *raison inverse* — treizième définition de l'élément cinquième — « comme si A est à B, ainsi que C à D, nous inférerons que par raison inverse, comme B est à A, ainsi D est à C ». La raison impliquée est alors la moitié[6].

### 1.2.2 Mathématiques élémentaires

En mathématiques élémentaires, les *tableaux de proportionnalité* constituent une application directe de la proportion analogique.

> Un tableau de nombres est un tableau de proportionnalité si l'on peut passer des nombres de la première ligne aux nombres correspondants de la seconde ligne en multipliant par un même nombre : le coefficient de proportionnalité.[7]

La raison, telle que définie par Euclide, est alors désignée par le terme de *coefficient de proportionnalité*, dont la nature dépend des quantités comparées. Il s'agit, par exemple, du taux de conversion entre pouces et centimètres dans le tableau 1.1.

---

[4] La remarque concernant les latins dans cette citation est sans aucun doute le fait du traducteur, Henrion (1632)

[5] 6 est le quart de 24, tout comme 3 est le quart de 12.

[6] 3 est la moitié de 6, tout comme 12 est la moitié de 24.

[7] Nous empruntons cette définition à un site d'assistance scolaire `http ://www.assistancescolaire.com`, consulté le 18 juillet 2014.

| | | |
|---|---|---|
| 3 | 5 | 12 |
| 7,62 | 12,7 | 30,48 |

TAB. 1.1 : Tableau de proportionnalité

Une seconde opération découle directement de l'existence de cette proportionnalité, il s'agit de la règle de trois, ou *règle de proportionnalité*, méthode permettant, dans une situation de proportionnalité connue, de déterminer la valeur d'une quantité inconnue. Par exemple, si nous disposons d'un corpus découpé en 1 226 124 formes, dont 236 847 noms communs, il est possible d'établir le pourcentage de noms communs du corpus à l'aide d'une règle de proportionnalité :

$$\frac{236\,847}{1\,226\,124} = \frac{x}{100} \Rightarrow \frac{236\,847 \times 100}{1\,226\,124} = x \Rightarrow x \simeq 19,32$$

Nous pouvons facilement traduire une telle opération sous la forme d'équation, appelée *équation analogique* ($236\,847 : 1\,226\,124 :: x : 100$) dont la résolution pourra être exprimée de la manière suivante : « 236 847 est à 1 226 124 ce que 19,32 est à 100 ». Bon nombre d'applications de l'analogie exploitent de telles équations, comme nous le verrons par la suite.

### 1.2.3 Analogies équivalentes

Lepage (2003) s'intéresse à l'analogie dans un contexte quelque peu différent de celui des mathématiques élémentaires. À la suite d'une étude détaillée de l'histoire de la notion, il propose en effet « de faire passer l'analogie entre chaînes de symboles d'un statut intuitif mais inutilisable automatiquement, à celui d'une opération désormais réapplicable aveuglément et donc aussi reproductible ». Il s'inscrit cependant dans une approche strictement proportionnelle, au sens d'Euclide et définit l'analogie comme « une conformité de rapports entre objets de même type » :

> une analogie fait toujours intervenir quatre objets, que nous avions décidé de noter A, B, C et D, et [...] l'analogie ou proportion est la conformité de leurs rapports, notée $A : B \doteq C : D$. Par définition, une analogie est une conformité de rapports entre objets de même type.
>
> Lepage (2003, p.111)

Ses observations l'amènent à formaliser un ensemble de propriétés, aboutissant à l'identification d'expressions analogiques équivalentes. Bien qu'il vérifie l'opérabilité de sa proposition sur les chaînes de symboles, les ensembles et les multi-ensembles, nous l'illustrons ici à l'aide de valeurs numériques.

#### 1.2.3.1 Cube des analogies équivalentes

Lepage (2003, 2006, 2014) propose de découper l'ensemble des analogies possibles pour quatre *objets* donnés (A, B, C et D) en trois classes de huit analogies équivalentes : $A : B \doteq C : D$, $A : C \doteq D : B$ et $A : D \doteq B : C$. Il s'appuie pour cela sur quelques propriétés de la *conformité* et des *rapports*.

La première de ces propriétés est la *réflexivité de la conformité*, selon laquelle « L'analogie $A : B \doteq A : B$ est vraie ». En effet, quel que soit le genre des objets A

et B, si nous les considérons en synchronie, le rapport qu'ils entretiennent est égal à lui-même. La deuxième propriété est la *symétrie de la conformité*, selon laquelle « si l'on peut dire que $A$ est à $B$ comme $C$ est à $D$, alors on doit pouvoir dire que $C$ est à $D$ comme $A$ est à $B$ ». Si nous reprenons l'exemple de proportion analogique entre longueurs de lignes $6 : 3 :: 24 : 12$, introduite en 1.2.1, nous pouvons effectivement déduire immédiatement l'analogie $24 : 12 :: 6 : 3$. Dans ce cas, le rapport considéré demeure identique, ce que 24 est à 12, 6 l'est à 3, son double.

Il s'appuie ensuite sur la propriété de *permutation des moyens*[8], qui concerne les rapports : « Soit l'analogie $A : B \doteq C : D$. Alors, de façon équivalente, l'analogie suivante est aussi vraie : $A : C \doteq B : D$ par permutation des moyens ». Cette propriété correspond à la permutation énoncée par Euclide, que nous avons introduite en 1.2.1. Rappelons que, dans ce cas, le rapport considéré varie, ce que 6 est à 24, 3 l'est à 12, son quart.

Lepage (2003, p.116) dérive alors deux autres propriétés de la symétrie de la conformité et de la permutation des moyens. Il montre que l'*inversion des rapports*, déjà introduite par Euclide, peut être décomposée en applications successives de ces deux propriétés :

$$\begin{array}{ll} A : B \doteq C : D & \\ \quad \Updownarrow & \text{(permutation des moyens)} \\ A : C \doteq B : D & \\ \quad \Updownarrow & \text{(symétrie de la conformité)} \\ B : D \doteq A : C & \\ \quad \Updownarrow & \text{(permutation des moyens)} \\ B : A \doteq D : C & \end{array}$$

Il en va de même pour la *permutation des extrêmes*, selon laquelle si A est à B comme C est à D, alors D est à B, ce que C est à A. Ainsi, dans le cas de notre exemple, $6 : 3 :: 24 : 12 \Leftrightarrow 12 : 3 :: 24 : 6$. Le rapport en œuvre dans l'analogie équivalente est alors différent du rapport initial, ce que 12 est à 3, 24 l'est à 6, son quadruple. La dérivation de cette propriété s'établit de la manière suivante :

$$\begin{array}{ll} A : B \doteq C : D & \\ \quad \Updownarrow & \text{(permutation des moyens)} \\ A : C \doteq B : D & \\ \quad \Updownarrow & \text{(symétrie de la conformité)} \\ B : D \doteq A : C & \\ \quad \Updownarrow & \text{(permutation des moyens)} \\ B : A \doteq D : C & \\ \quad \Updownarrow & \text{(symétrie de la conformité)} \\ D : C \doteq B : A & \\ \quad \Updownarrow & \text{(permutation des moyens)} \\ D : B \doteq C : A & \end{array}$$

Lepage dispose alors de trois propriétés permettant de représenter chaque classe de huit analogies équivalentes sous forme de cube. La figure 1.1 présente la classe

[8]Le terme de *permutation des moyens* est hérité des travaux de Hermann (1960).

correspondant à l'analogie $A : B :: C : D$. La symétrie de la conformité y est représentée par les lignes formées de tirets longs, l'inversion des rapports par les lignes pointillées et la permutation des extrêmes par les lignes continues.

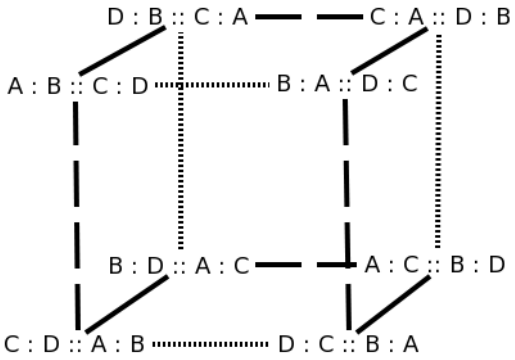



FIG. 1.1 : Cube des analogies équivalentes à $A : B :: C : D$

Comme nous l'avons vu au fil de cette section, l'ensemble des analogies équivalentes varie en termes de rapport. La figure 1.2 reprend l'exemple des longueurs de lignes qui nous a servi jusqu'à présent. À gauche, elle représente l'ensemble des analogies équivalentes correspondant à la classe de l'analogie « 6 est à 3 ce que 24 est à 12 ». À droite, nous avons remplacé chaque analogie par le rapport qu'elle amène à énoncer, « 6 est à 3 ce que 24 est à 12, son **double** ». Cette représentation met en évidence le fait que la proportion analogique n'est pas basée sur la conformité d'un seul rapport, mais de plusieurs.

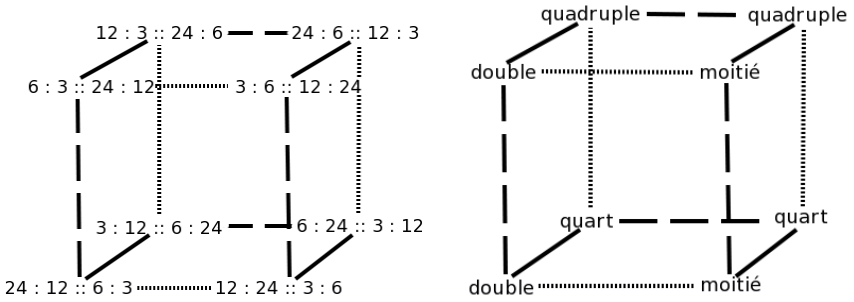



FIG. 1.2 : Cube des analogies équivalentes à $6 : 3 :: 24 : 12$ et rapports

#### 1.2.3.2 Similarité et contiguïté des objets

Après avoir exploité les propriétés de la conformité et des rapports, Lepage (2003) cherche à formaliser l'analogie relativement aux propriétés des *objets* qu'elle met en relation. Il s'intéresse alors à deux propriétés révélées par son étude historique, la *contiguïté* et la *similarité*.

**Contiguïté**

Lepage évoque à plusieurs reprises les difficultés rencontrées avec la notion de contiguïté, dont il ne parvient pas à établir d'expression mathématique exploitable dans le cadre des analogies entre chaînes de symboles. Il définit cette propriété comme une relation entre les objets faisant intervenir l'existence d'un tout.

> La notion de contiguïté caractérise la relation de la partie au tout, du tout à la partie ou de la conjonction de deux parties dans l'expérience [...] la contiguïté devrait donc caractériser le fait qu'une sous-chaîne (ou un symbole) appartienne à une sous-chaîne, ou qu'une sous-chaîne contienne une sous-chaîne (ou un symbole), ou que deux sous-chaînes se suivent (connectivité).
>
> Lepage (2003, p.103)

Il s'appuie sur celle-ci pour postuler l'existence d'une équivalence supplémentaire entre analogies, qu'il nomme *inversion des objets* :

$$A : B \doteq C : D \Leftrightarrow A^{-1} : B^{-1} \doteq C^{-1} : D^{-1}$$

L'inversion des objets sous-entend l'existence d'un ensemble de référence, accessible à travers les seuls quatre objets A, B, C et D et dans lequel il serait possible d'atteindre « l'inverse le plus contigu » de chacun des objets. Il ne parvient cependant à matérialiser cette propriété que dans le cas des proportions d'ensembles. Il considère alors l'ensemble de référence comme l'union des quatre objets et l'inverse de chacun d'entre eux comme son complémentaire.

**Similarité**

La notion de similarité est plus simple à formaliser. Pour que des objets puissent être mis en relation de conformité de rapport, ils doivent partager quelque chose de commun. C'est ce qu'Euclide énonçait par « quantités de même genre ». Lepage va plus loin dans son analyse. Il parle d'une « dimension selon laquelle les rapports seront établis » et considère que la similarité entre objets au regard de cette dimension est entièrement consommée par la relation d'analogie. Ainsi, lorsque nous énoncons que A est à B ce que C est à D, nous focalisons notre attention sur certaines propriétés de A, que cet objet partage entièrement, soit avec B, soit avec C. Ce que Lepage formalise de la manière suivante :

> **(Distribution)** *Soit l'analogie $A : B \doteq C : D$, toute propriété de A se retrouve dans B ou dans C.*
>
> Lepage (2003, p.122)

De la même façon que les propriétés partagées par A et B seront partagées par C et D, celles partagées par A et C le seront par B et D. En parvenant à établir une mesure de similarité entre objets, il sera donc possible de résoudre les équations analogiques. Dans le cas des chaînes de symboles, il est possible de s'appuyer sur les distances d'édition pour effectuer cette mesure. Nous reviendrons sur ce point dans le chapitre 3, section 3.1.1.

Lepage (2003, p.103,106) regrette de ne pas être parvenu à faire émerger une notion unique, englobant les deux dimensions de similarité et de contiguïté, comme le suggèrent les ensembles d'analogies équivalentes.

### 1.2.4 Conséquences

Nous avons vu dans cette première partie que l'analogie peut-être considérée sous un angle mathématique, comme une conformité mettant en œuvre quatre objets entretenant des relations qu'il est possible de quantifier et de manipuler automatiquement. La similarité entre ces objets semble être la clef de la constitution de proportions analogiques, puisque c'est elle qui permet d'énoncer les rapports. Elle dépend entièrement de la nature des objets, question à laquelle il nous faudra porter une attention particulière. Nous avons également vu qu'une analogie énoncée pour mettre en évidence un rapport particulier sous-entend l'existence d'autres rapports.

Reprenons l'exemple donné en introduction d'une analogie entre *aboyer* et *beugler*. Admettons que l'action de pousser un cri pour un certain animal correspond à leurs sens premiers (ABOYER I et BEUGLER I) et la façon particulière de s'exprimer à leurs sens seconds (ABOYER II et BEUGLER II). Si nous établissons la proportion analogique ABOYER II : ABOYER I :: BEUGLER II : BEUGLER I, en nous basant sur une conformité de rapport de type « le second sens est un sens métaphorique du premier », nous énonçons également une conformité de rapport entre ABOYER II et BEUGLER II d'une part et ABOYER I et BEUGLER I d'autre part. Si cette conformité est absente, l'analogie n'est pas valide.

## 1.3 Raisonnement analogique

La philosophie s'intéresse très tôt à la transposition de l'analogie mathématique à d'autres domaines de réflexions, notamment la métaphysique. Lepage (2003) cite Olympiodore le Jeune (VI$^{e}$ s.), dans les travaux de qui l'analogie est considérée comme un mode d'enseignement, un moyen de persuasion, une figure de style. Il cite également Aristote (IV$^{e}$ s. av. J.-C.), qui considère l'analogie comme un cas particulier de métaphore d'une part et un outil de raisonnement d'autre part. Dans la présente section, nous nous concentrons sur l'analogie en tant qu'outil de raisonnement et sur les travaux de Gentner qui en propose une formalisation. Nous commencerons par définir le procédé en lui-même, avant de présenter la terminologie utilisée pour l'appréhender. Nous évoquerons ensuite la place de l'analogie dans les différents types de comparaison avant de réaliser un parallèle entre cette approche et les proportions.

### 1.3.1 Procédé de compréhension

Gentner (1983), Falkenhainer et al. (1989), Gentner et Holyoak (1997) et Gentner et Markman (1997), dont les travaux se situent en sciences cognitives, s'intéressent à l'analogie en tant que procédé de compréhension d'une nouvelle situation par exploitation des connaissances acquises dans une situation préalable. Un mécanisme cognitif qui permet d'effectuer des inférences d'un domaine à un autre, d'apprendre des catégories et des schémas englobant deux situations comparées.

> But how do categories first get formed? One basic mechanism is *analogy* — the process of understanding a novel situation in terms of one that is already familiar.[...] And the analogy between two specific situations

> may provide the "seed" for learning a more general category or schema that encompasses both.
>
> Gentner et Holyoak (1997, p.32)

Gentner (1983) parle alors de transfert de *connaissances* d'un domaine ou d'une situation *source*, connu, vers un domaine ou une situation *cible*, moins bien maîtrisé. Les analogies qu'elle énonce ne mettent en œuvre que deux éléments : « Une batterie électrique est comme un réservoir »[9], « L'atome d'hydrogène est comme notre système solaire »[10], « la chaleur est comme l'eau »[11]. Cependant elle propose de formaliser ce mode de comparaison d'une façon qui laisse entièrement leur place aux proportions analogiques, même si elle ne l'exprime pas en ces termes.

Cet aspect de l'analogie joue un rôle dans les découvertes scientifiques. Gentner (1983) et Gentner et Markman (1997) font référence aux analogies de Kepler concernant les phénomènes astronomiques et à celle de Rutherford, s'appuyant sur ses connaissances du système solaire pour mieux comprendre le fonctionnement de l'atome. Bontems (2007), pour sa part, nous apprend l'importance du *principe d'analogicité* dans les travaux de Ferdinand Gonseth[12], qu'il situe entre les modèles abstraits et les situations concrètes.

> Cette correspondance entre le rationnel et le réel est progressivement identifiée comme étant un principe fondateur de toute connaissance : le « principe d'analogicité ». Il y a un autre principe fondamental de ce type, le « principe de causalité »[...].
>
> La science consiste en systèmes de relations causales construits par l'esprit et soumis à l'expérience afin d'établir leur correspondance analogique avec la réalité extérieure.
>
> Bontems (2007, p.5)

Les sciences de l'éducation s'intéressent également au raisonnement analogique, en tant qu'outil pédagogique. Ainsi, Wong (1993) développe l'idée selon laquelle la construction d'analogies personnelles permet à un apprenant de s'approprier de nouveaux concepts scientifiques en les situant par rapport à ses propres connaissances et qu'elle favorise l'abstraction.

> Constructing one's own analogy serves to (a) make new situations familiar, (b) represent the problem in the particulars of individual's prior knowledge, and (c) stimulate abstract thinking about underlying structure or patterns.
>
> Wong (1993, p.377)

Gentner et Markman (1997), ainsi que Wong (1993) évoquent la possibilité d'effectuer de mauvaises inférences par raisonnement analogique. L'analogie ne doit pas être considérée comme une preuve scientifique, mais plutôt comme un outil exploratoire. Cette considération rejoint celles d'Aristote relevées par Lloyd (1966).

[9] *An electric battery is like a reservoir.*
[10] *The hydrogene atom is like our solar system.*
[11] *Heat is like water.*
[12] Ferdinand Gonseth [1890-1975] est un épistémologue suisse, mathématicien de formation.

### 1.3.2 Système d'objets

La formalisation du raisonnement analogique proposée par Gentner (1983) s'appuie sur une description détaillée des domaines, ou situations, confrontés l'un à l'autre. Dans ce cadre, un domaine est vu comme un système d'*objets*, lesquels disposent d'*Attributs* et entretiennent entre eux des *Relations*. Les connaissances, pour leur part, sont vues comme des réseaux de prédicats. Un prédicat à un seul argument est considéré comme un Attribut de l'objet auquel il se rattache. Un argument à plus d'un argument est considéré comme une Relation. Elle distingue également les prédicats de premier ordre, qui ont pour arguments un ou des objets, des prédicats du second ordre, qui ont pour arguments des prédicats.

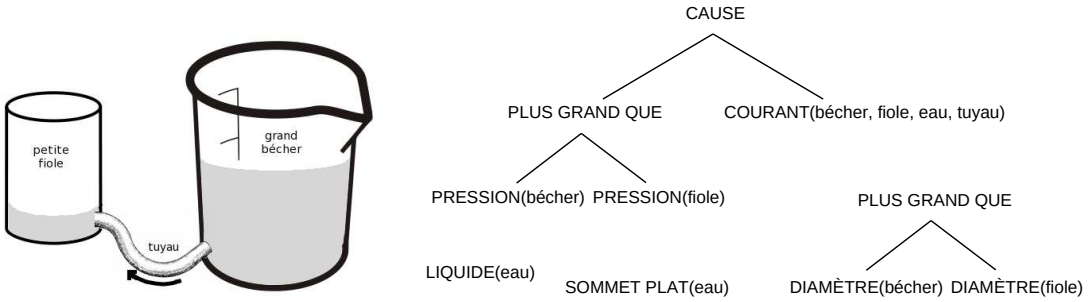



FIG. 1.3 : Situation physique de courant d'eau et description

La figure 1.3 regroupe deux figures empruntées à Falkenhainer et al. (1989). Elle fournit un exemple de description, simplifiée, construite de cette manière. La partie gauche illustre la situation physique de déplacement d'eau d'un grand bécher à une petite fiole, par l'intermédiaire d'un tuyau. À droite, cette situation est décrite à l'aide de quatre objets (bécher, fiole, eau, tuyau) et de dix prédicats. Les prédicats se répartissent en quatre Attributs, tels que DIAMÈTRE(fiole), une Relation entre objets, COURANT(bécher, fiole, eau, tuyau), et trois Relations d'ordre supérieur, telles que PLUS GRAND QUE [DIAMÈTRE(bécher), DIAMÈTRE(fiole)]. Parmi eux, le prédicat d'ordre supérieur CAUSE(PLUS GRAND QUE [PRESSION(bécher), PRESSION(fiole)], COURANT(bécher, fiole, eau, tuyau)) structure un sous-système de Relations interdépendantes.

### 1.3.3 Type de comparaison

Gentner oppose l'analogie à deux autres types de comparaisons : la *similarité littérale* et l'*abstraction*. Lors d'une comparaison par similarité littérale, la majorité des prédicats de la source sont mis en correspondance avec les prédicats de la cible, indépendamment du nombre d'arguments qu'ils mettent en jeu, comme dans l'exemple « Le système stellaire X12, dans la galaxie d'Andromède est comme notre système solaire. »[13]. Lors d'une comparaison par analogie, en revanche, les prédicats relationnels sont transposés en priorité, la conservation des Attributs des objets entre le domaine source et le domaine cible étant de moindre importance que les Relations qu'entretiennent les objets. La différence entre ces deux types de comparaison n'est cependant pas une opposition radicale, mais plutôt un continuum.

[13] *The X12 star system in the Andromeda galaxy is like our solar system.*

Le dernier type de comparaison, l'abstraction, se différencie de l'analogie par la nature des objets en jeu dans le domaine source. Dans le cas d'une abstraction, il s'agit de concepts, comme dans l'exemple « L'atome d'hydrogène est un système à force centrale. »[14], tandis que dans le cas de l'analogie, il s'agit d'objets concrets. De plus, dans le cas d'une abstraction, l'ensemble des prédicats de la source est transféré sur la cible, ce qui n'est pas nécessairement le cas pour l'analogie.

Ces trois cas de comparaison amènent à effectuer des *inférences*. La situation source étant mieux connue que la situation cible, sa description sera plus détaillée et comportera notamment davantage de prédicats d'ordre supérieur. Une fois le transfert des connaissances effectué, certains éléments de la situation cible n'auront pas trouvé de correspondance et pourront être proposés comme candidat à l'inférence. Dans le cas de l'analogie, toutes les Relations non transférées ne seront cependant pas candidates. Gentner s'intéresse à cette question et cherche à déterminer un critère de priorité dans le transfert des prédicats. Elle observe une préférence pour le transfert de structures de prédicats interconnectés sur celui de prédicats isolés. Elle nomme cela le *principe de systématicité*. Les prédicats d'ordre supérieur joueraient un rôle important, permettant de structurer la connaissance à l'aide de Relations causales, mathématiques ou fonctionnelles. Ils participeraient également à contrôler la cohérence. Ainsi, dans le cas d'une analogie dont la source correspondrait à la description présentée dans la figure 1.3, c'est le système de Relations interdépendantes CAUSE(...) qui serait transféré en priorité[15].

Contrairement à ce que nous avons vu pour l'analogie proportionnelle, l'analogie comme type de raisonnement n'est pas nécessairement symétrique. Selon Gentner et Markman (1997), il existe une préférence cognitive à effectuer la comparaison par analogie d'une source plus cohérente, en terme de structure, vers une cible moins cohérente. De la sorte, davantage d'inférences pourront être effectuées et l'analogie s'avérera plus informative.

### 1.3.4 Raisonnement analogique et proportions

Les exemples de situation physique et de description de celle-ci empruntés à Falkenhainer et al. (1989) dans la figure 1.3, page suivante, sont initialement associés à une seconde situation physique et à sa description, pour illustrer l'analogie « la chaleur est comme l'eau ». La figure 1.4 reprend ces deux éléments.

Sans reprendre en détail l'appariement effectué entre ces deux situations, nous nous intéressons ici à son résultat. Comme nous l'avons déjà mentionné, d'après le principe de systématicité, les Relations isolées sont laissées de côté et seul le système de Relations interdépendantes gouverné par la Relation de causalité dans la description du courant d'eau est transféré vers la description du courant de chaleur. Les objets des deux situations sont mis en correspondance deux à deux : bécher → café, fiole → glaçon, eau → chaleur, tuyau → barre en argent. Le prédicat PLUS GRAND QUE [PRESSION(bécher), PRESSION(fiole)] est associé au prédi-

[14] *The hydrogen atom is a central force system.*

[15] Ce principe de systématicité, établi lors d'une étude de l'analogie, a été étendu par la suite à la comparaison par similarité littérale.

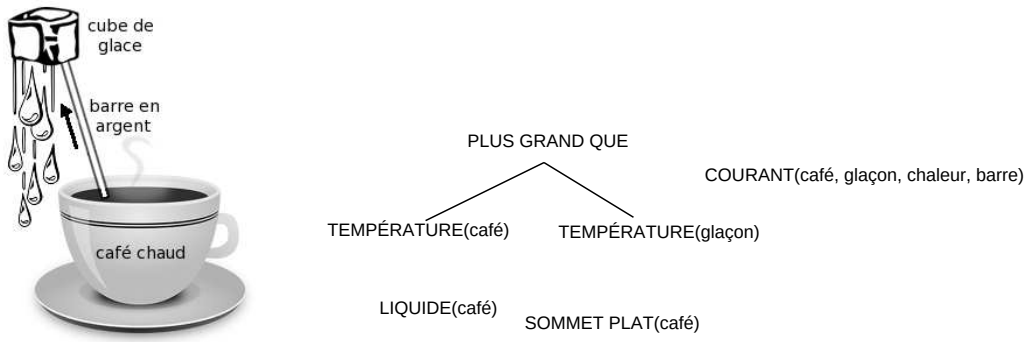



FIG. 1.4 : Situation physique de courant de chaleur et description

cat PLUS GRAND QUE [TEMPÉRATURE(café), TEMPÉRATURE(glaçon)] et le prédicat COURANT(bécher, fiole, eau, tuyau) au prédicat COURANT(café, glaçon, chaleur, barre). Le prédicat de causalité, absent de la description du courant de chaleur, est inféré.

Le résultat de cet appariement peut être énoncé sous la forme d'une liste d'analogies proportionnelles :

- « La pression du bécher **est à** la pression de la fiole **ce que** la température du café **est à** la température du glaçon. »
- « Le rapport de grandeur entre la pression du bécher et de la fiole **est au** courant entre le bécher, la fiole, l'eau et le tuyau **ce que** le rapport de grandeur entre la température du café et la température du glaçon **est au** courant entre le café, le glaçon, la chaleur et la barre en argent. »
- « Le bécher **est à** la fiole **ce que** le café **est au** glaçon. »

La rectification et l'extension des analogies, évoquées par Wong (1993) et Gentner et Markman (1997) comme un moyen de raisonnement plus économique que la création de nouvelles analogies, consiste alors à réviser les proportions analogiques sous-entendues par l'appariement. Cette vision de l'analogie mettant en jeu des proportions entre domaines est proche de la proposition de Nagao (1984) pour la traduction automatique entre japonais et anglais. Elle est également conforme à la méthode proposée par Turney (2008a) pour la résolution d'analogies scientifiques et de métaphores courantes. Elle est cependant rejetée par Lepage (2003), qui considère la conformité de rapports entre objets de domaines différents difficile à établir.

### 1.3.5 Homomorphisme entre domaines

Rappelons-le, dans le cadre de la formalisation proposée par Lepage (2003), rien n'est nécessaire pour interpréter une analogie que les quatre objets qu'elle met en relation. De plus, la conformité de rapport dont elle rend compte se base sur une similarité entre objets, au regard d'une unique dimension, qui est entièrement consommée par l'analogie. Une analogie de type *je donne* : *tu donnes* :: donner_ind_pres_1_sg :

donner_ind_pres_2_sg, qui serait autorisée par le modèle de Nagao (1984), pose donc problème[16]. La similarité en œuvre dans le couple (*je donne*, *tu donnes*) n'est pas établie selon la même dimension que celle en œuvre dans le couple (*je donne*, donner_ind_pres_1_sg). Tandis que la première est d'ordre graphique, la seconde est de l'ordre de la mise en correspondance entre objets de domaines distincts. Il ne s'agit donc pas à proprement parler d'une comparaison.

Lepage (2003) avance l'idée que la mise en correspondance d'objets de deux domaines n'est possible que si ces objets vérifient des analogies, chacun dans leur domaine. Ainsi, la mise en correspondance (*je donne*, donner_ind_pres_1_sg) sous-entendrait, a minima, l'existence d'une analogie $je\ donne : x :: y : z$ dans le domaine des formes verbales et $\mathsf{donner\_ind\_pres\_1\_sg} : a :: b : c$ dans celui des descriptions d'analyses flexionnelles. Et c'est seulement si les structures induites par ces analogies sont identiques que la mise en correspondance des domaines est possible. C'est donc l'*homomorphisme*, en tant qu'« application d'un ensemble muni d'une loi de composition interne dans un autre ensemble muni d'une autre loi de composition interne » qui la rend possible.

Cette vision des choses permet à Lepage (2003) d'aborder la question de l'intégration parallèle d'une nouvelle observation dans deux domaines — dont certains objets ont préalablement été mis en correspondance — sous l'angle de la résolution d'équations analogiques internes à chacun. Il applique cette méthode à l'analyse syntaxique et à la traduction automatique.

### 1.3.6 Conséquences

Nous avons vu dans cette section que l'analogie peut-être considérée comme un processus de comparaison de deux domaines structurés en système de relations. Comme le soulignent Falkenhainer et al. (1989) et Gentner et Markman (1997), l'automatisation de cette comparaison nécessite une représentation des domaines et des relations suffisamment explicites pour pouvoir rendre compte des dépendances. Veale et Li (2014) soulignent que si cette approche a abouti à la mise en œuvre de traitements informatiques très différents de l'approche par proportions, ces traitements sont à présent rendus compatibles. Cette compatibilité est assurée par l'idée, déjà avancée par Turney (2008b), que les systèmes analogiques complexes peuvent être vus comme des combinaisons de proportions analogiques plus petites et plus locales.

Reprenons l'exemple donné en introduction d'une analogie entre ABOYER et BEUGLER. Nous avons mis l'accent, dans la partie précédente, sur un rapport particulier existant entre les deux sens de chacun de ces vocables[17], la métaphore. Nous aurions pu tout aussi bien nous concentrer sur le sens premier de chacun d'entre eux et le rapport qu'ils entretiennent avec l'animal dont ils dénotent l'action de pousser le cri. Ainsi, nous aurions pu énoncer que « ABOYER❙ est à CHIEN❙ ce que BEUGLER❙ est à BŒUF❙ »[18]. Sans pousser plus loin l'exploration des analogies proportionnelles

[16]Les exemples de conjugaison des verbes français utilisés ici sont adaptés de Lepage (2003, pp. 291-293).

[17]La notion de vocable sera présentée dans le chapitre 2, section 2.1.2.

[18]Nous supposons ici que les sens premiers de CHIEN et de BŒUF, numérotés ❙, correspondent

sous-jacentes à l'énoncé « ABOYER est analogue à BEUGLER », nous pouvons déjà émettre l'hypothèse qu'elle n'est pas basée sur une unique proportion analogique mettant en œuvre quatre unités lexicales, mais sur un ensemble interdépendant plus large. Nous supposons alors qu'une ressource dictionnairique rendant compte de façon explicite des relations qu'entretiennent les unités lexicales permettra au locuteur d'affiner son sentiment de similitude entre mots et nous permettra d'acquérir automatiquement des connaissances sur l'organisation du lexique, par abstraction, à partir des analogies qu'elle contient.

## 1.4 Analogie formelle

Avant de nous intéresser à un certain nombre de travaux relatifs à l'étude du lexique et exploitant l'analogie, nous souhaitons présenter le cas particulier de la transposition de l'analogie mathématique aux chaînes de symboles. De telles analogies, que nous appellerons désormais *analogies formelles*, sont considérées par Saussure (1916) comme un mécanisme linguistique de création et d'évolution lexicales. Elles ont été formalisées par Lepage (2003), Stroppa et Yvon (2005) et Stroppa (2005).

### 1.4.1 Mécanisme de création

Saussure (1916) présente l'analogie comme un procédé de création de formes nouvelles dans une langue donnée à partir de formes préexistant dans cette même langue. Ses considérations s'inscrivent donc dans le cadre d'une étude de l'analogie formelle. Il rapporte ce procédé à la résolution d'équations analogiques à l'aide de règles de proportionnalité, telle que nous les avons présentées dans la section 1.2.1.

Il insiste à la fois sur la simplicité de la mise en œuvre de cette analogie et sur le caractère régulier des formes qui en découlent, basées sur l'exploitation d'« un rapport unissant les formes entre elles ».

> Pour former *indécorable*, nul besoin d'en extraire les éléments (*in-décorable*) ; il suffit de prendre l'ensemble et de le placer dans l'équation :
>
> *pardonner : impardonnable*, etc., = *décorer : x*.
> *x = indécorable*.
>
> De la sorte on ne suppose pas chez le sujet une opération compliquée, trop semblable à l'analyse consciente du grammairien.
>
> Saussure (1916)

Il évoque également la modification de formes par analogie. Selon lui, la création d'une forme nouvelle entrant en concurrence avec une forme préexistante ne vient cependant pas *remplacer* celle-ci. La disparition de la forme première, moins régulière, demeure optionnelle et est indépendante du procédé de création analogique.

---

au sens d'animal qui nous intéresse.

Néanmoins, l'analogie joue selon lui un rôle dans le maintien des formes existantes. Il distingue alors deux types de formes dont le maintien serait assuré dans une langue donnée : des mots isolés, tels que les noms propres, pour lesquels il n'envisage aucune innovation lexicale concurrente possible et des formes étroitement liées aux formes de leur entourage lexical.

> Ainsi les formes se maintiennent parce qu'elles sont sans cesse refaites analogiquement ; un mot est compris à la fois comme unité et comme syntagme, et il est conservé pour autant que ses éléments ne changent pas. Inversement son existence n'est compromise que dans la mesure où ses éléments sortent de l'usage. Voyez ce qui se passe en français pour *dites* et *faites* [...] la langue cherche à les remplacer ; on entend dire *disez*, *faisez*, sur le modèle de *plaisez*, *lisez*, etc., et ces nouvelles finales sont déjà usuelles dans la plupart des composés (*contredisez*, etc.).
>
> Saussure (1916)

Paralèllement à ces observations, il souligne l'importance de la potentialité des langues révélée par la production analogique de formes nouvelles. Chacune de ces formes « existe déjà en puissance dans la langue » et c'est cette possibilité qui, selon lui, est le fait le plus remarquable.

### 1.4.2 Formalisation

La formalisation proposée par Lepage (2003), Stroppa et Yvon (2005) et Stroppa (2005) s'applique aux analogies formelles entre chaînes de caractères étudiées par Saussure (1916), tout autant qu'à celles mettant en jeu des symboles de différentes natures : structures de traits, transcriptions phonétiques, images, etc. Les rapports considérés entre ces chaînes de symboles sont d'ordre graphique et non sémantique. De telles analogies sont cependant souvent exploitées pour relever des *analogies sémantiques*, considérées en traitement automatique des langues comme de « vraies » analogies.

L'analogie *engager* : *désengagement* :: *abonner* : *désabonnement* peut être traitée comme une analogie formelle. Les différences graphémiques entre *engager* et *désengagement* sont identiques à celles entre *abonner* et *désabonnement*.

| $\epsilon$ | **engag** | er | $\epsilon$ | **abonn** | er |
|---|---|---|---|---|---|
| dés | **engag** | ement | dés | **abonn** | ement |

TAB. 1.2 : (*engager*,*désengagement*) vs. (*abonner*,*désabonnement*)

Comme nous le voyons dans le tableau 1.2, qui rend compte de la régularité de différences énoncée plus haut, il est possible de considérer chaque chaîne de symboles comme une suite de sous-chaînes. Stroppa (2005) et Stroppa et Yvon (2005) parlent alors de *factorisation*. Chaque mot peut être décomposé en *facteurs*, auxquels il est possible d'associer des indices, par ordre d'apparition. Une fois cette factorisation réalisée, la validation de l'analogie passe par la vérification d'une alternance entre

facteurs deux à deux.

Étant donnée l'analogie $a : b :: c : d$, pour un indice donné, soit les facteurs de $a$ et $b$ sont identiques, tout autant que les facteurs de $c$ et $d$, soit les facteurs de $a$ et $c$ sont identiques, tout autant que les facteurs de $b$ et $d$. Sinon, l'analogie formelle n'est pas valide.

Dans le cas de *engager* : *désengagement* :: *abonner* : *désabonnement*, les factorisations mettant en jeu le moins de facteurs possibles sont les suivantes :

- $a =$ engager ; $f_a = (\epsilon, engag, er)$
- $b =$ désengagement ; $f_b = (dés, engag, ement)$
- $c =$ abonner ; $f_c = (\epsilon, abonn, er)$
- $d =$ désabonner ; $f_d = (dés, abonn, ement)$

L'alternance de facteurs est alors bien vérifiée : ($\epsilon$, *dés*, $\epsilon$, *dés*) pour les facteurs d'indice 1, (*engag*, *engag*, *abonn*, *abonn*) pour ceux d'indice 2 et (*er*, *ement*, *er*, *ement*) pour ceux d'indice 3.

Lepage, Stroppa et Yvon exploitent cette formalisation de l'analogie formelle dans différentes tâches de traitement automatique des langues, parmi lesquelles l'automatisation d'analyses linguistiques (Pirrelli et Yvon, 1999 ; Stroppa, 2005 ; Stroppa et Yvon, 2006), la traduction automatique (Lepage, 2006 ; Takeya et al., 2011) et l'analyse de caractères chinois (Lepage, 2014).

## 1.5 Lexique et analogie

### 1.5.1 Réhabilitation de l'analogie

Lepage (2003) et Dal (2003) soulignent tous deux l'importance accordée à l'analogie dans l'étude de l'évolution du lexique jusqu'à la moitié du XX$^{e}$ siècle et à la prépondérance du générativisme. Dal (2003) explique le rejet de l'analogie chez les générativistes par deux raisons principales : le déplacement du point de vue sur la langue de celui du locuteur vers celui du linguiste et l'incompatibilité entre la prise en compte de l'existant et les règles de grammaire générative transformationnelle. Elle souligne l'aspect pragmatique et surfacique de l'analogie, qui s'intéresse directement au mot et non aux morphèmes. Lepage (2001), pour sa part, s'intéresse à trois arguments précis avancés par les générativistes contre l'analogie : l'hypothèse de l'inné, l'hypothèse du hors-contexte et la surproduction. Il confronte sa formalisation de l'analogie à ses trois arguments et montre ainsi qu'il est possible de dépasser ces critiques.

Après une discussion des facteurs permettant la réhabilitation de l'analogie au début du XXI$^{e}$ siècle, Dal (2003) s'intéresse aux innovations lexicales observables en corpus. Elle souligne que bien souvent, il est difficile de déterminer si ceux-ci sont le fruit de l'application de règles ordonnées ou d'un procédé analogique.

> La plupart du temps en effet, aucun argument décisif ne permet de dire si elles sont le produit de l'application de règles ordonnées à partir de matériau en entrée, ou si c'est par analogie avec des mots existants qu'elles ont été produites. Par exemple, le même nom *choucrouterie*, relevé dans les archives de *Libération* [...], mais absent des principaux dictionnaires synchroniques actuels, peut être utilisé par les partisans d'une morphologie basée sur règles, ou par ceux d'une morphologie basée sur l'analogie : les premiers diront qu'il illustre l'aptitude de *-erie* à s'appliquer à des noms d'artefacts pour donner des noms d'activité [...] ; les seconds verront en lui la solution de l'équation *croissant*:*croissanterie* = *choucroute*:*x*, sans qu'ils jugent nécessaire de le décomposer par ailleurs en *choucroute -erie*.
>
> Dal (2003, p.15)

Elle s'intéresse alors à un certain nombre de cas pour lesquels une approche à base de règles ordonnées ne semble pas aboutir à des propositions satisfaisantes. Elle montre que l'approche à base d'analogie propose, pour ces cas, des solutions moins coûteuses et plus proches de l'intuition. Par exemple, dans le cas de la substitution de suffixes entre les noms en *-teur* et les noms en *-trice*, tels qu'*acteur* et *actrice*, l'ordonnancement de règles nécessite la mise en place d'une règle de troncation de *-teur* sur la forme masculine, puis d'une règle de concaténation de *-trice*. Le recours à une équation analogique de type *acteur*:*actrice* = *sénateur*:*x* pour formaliser la construction de *sénatrice* est, elle, plus directe.

Elle souligne également que l'analyse de néologismes formels en corpus amène à observer de nombreux cas où ils cooccurrent avec des mots attestés mettant en œuvre le même procédé de construction. Ce qui conforte l'idée selon laquelle les locuteurs auraient recours à l'analogie pour produire de nouveaux mots.

Lepage (2001) et Dal (2003) considèrent l'un comme l'autre que l'analogie n'est pas seule en œuvre dans la langue et dans la construction du lexique. Dal (2003) souligne toutefois qu'elle peut être une façon intéressante d'appréhender les quantités de données attestées actuellement disponibles et qu'« il reste au morphologue à décrire les régularités qu'il perçoit dans le lexique construit, peu important finalement que, d'un point de vue cognitif, ces régularités soient abstraites par les locuteurs des mots complexes qu'ils connaissent sous la forme de règles symboliques, ou qu'elles soient incarnées dans des instances particulières ».

### 1.5.2 Morphologie dérivationnelle

Hathout (2009, 2011) s'inscrit dans la lignée de la proposition de Dal (2003) d'exploiter l'analogie pour étudier la construction lexicale. Il s'intéresse à l'analyse morphologique dérivationnelle « dans le cadre d'une théorie morphologique lexématique dans laquelle les atomes ne sont pas des morphèmes mais des mots ». Plus précisément, il cherche à constituer automatiquement une ressource lexicale, Morphonette[19], rendant compte de l'existence des familles morphologiques et des séries dérivationnelles qui structurent le lexique du français. Il utilise pour matériau de

[19] `http ://redac.univ-tlse2.fr/lexiques/morphonette.html`

base une ressource dictionnairique, le TLFi[20].

Par familles morphologiques, il entend des ensembles de mots qui partagent des propriétés formelles et sémantiques les plus spécifiques possible. Par séries, il entend des ensembles de mots les plus larges possible qui partagent des propriétés sémantiques et formelles générales et participent au plus grand nombre possible d'analogies qui impliquent les autres membres de la série. Par exemple, « *modifiable, modificateur, modifier, modification* » forment une famille morphologique dérivationnelle, tandis que « *modifiable, fructifiable, rectifiable, sanctifiable* » forment une série dérivationnelle.

La détection d'analogies, formelles et sémantiques, est donc au cœur de ses préoccupations. La méthodologie qu'il développe est constituée de trois étapes. La première consiste à délimiter le champ des analogies possibles à l'aide d'une mesure de *similarité morphologique*. Cette mesure est effectuée sur un bi-graphe et exploite la méthode de balade aléatoire de Gaume (2004). Un côté du graphe est constitué de l'ensemble des entrées du TLFi, le second côté de traits formels et/ou sémantiques[21]. Pour chaque entrée, un ensemble de *plus proches voisins* est calculé de manière à ce que plus deux entrées partagent de traits, plus elles soient considérées comme étant des voisins proches. Chaque ensemble ainsi obtenu contient tout à la fois des membres de la famille morphologique dérivationnelle de l'entrée, des membres de sa série dérivationnelle et de mauvais candidats n'appartenant à aucune de ces deux catégories. La figure 1.5 reprend l'ensemble des 50 plus proches voisins obtenus par Hathout (2011) pour *fructifier*. La mise en forme du texte y est celle d'origine : les membres de la famille de *fructifier* apparaissent en gras, ceux de sa série en italique et les mauvais candidats sont soulignés.

**fructifier fructifiant fructificateur fructification fructifère** *sanctifier rectifier présanctifier* **fructivore** fructidorien fructidorienne fructidoriser fructidor **fructueusement fructueux fructuosité fructose** obstructif constructif instructif désobstructif destructif autodestructif **usufructaire infructueusement infructueux infructuosité** sanctifiant sanctifiable rectifieuse rectifieur rectifiant rectifiable *transsubstantifier substantifier stratifier cimentifier certifier savantifier refortifier ratifier présentifier pontifier plastifier notifier nettifier mortifier mythifier mystifier quantifier*

FIG. 1.5 : 50 plus proches voisins de *fructifier*

La seconde étape du traitement, consiste à la vérification de l'ensemble des proportions analogiques $a : b :: c : d$ et $a : c :: b : d$ possibles correspondant au modèle suivant : $b$ est un voisin de $a$, $c$ est un autre voisin de $a$ plus éloigné que $b$ et $d$ est à la fois un voisin de $b$ et de $c$.

[20] `http ://atilf.atilf.fr/`

[21] Hathout (2009) compare l'utilisation de $n$-grammes de lettres, comme traits formels, et de $n$-grammes de mots apparaissant dans les définitions du TLFi, comme traits sémantiques. Dans Hathout (2011), seuls des traits formels sont exploités à cette étape, des $n$-grammes de transcriptions phonétiques.

Les proportions analogiques qu'il constitue ainsi sont des analogies formelles, telles que nous les avons présentées en 1.4. Dans un premier temps, Hathout (2009) utilise la factorisation de Stroppa et Yvon (2005). Par la suite, pour diminuer la complexité du traitement au risque de perdre en exhaustivité, Hathout (2011) a recours à une méthode alternative. Il associe à chaque couple de mots une signature, correspondant à une transformation de type distance de Levenshtein. Il cite en exemple le couple (*fructueux, infructueusement*), qui a pour signature : (insert, $\epsilon$, *in*), @, (replace,$x$,*sement*). Nous retrouvons dans cette signature la notion de factorisation, *fructueux* et *infructueusement* comptent chacun trois facteurs. Pour passer du premier mot au second, *in* est inséré à la place du facteur vide, le second facteur est conservé et le dernier est remplacé par *sement*.

Les membres de chacun des couples pour lesquels la relation analogique est vérifiée sont considérés comme appartenant à une même famille ou série. Ils sont alors stockés comme des sommets d'un graphe, reliés par une relation morphologique. Une première estimation de la distinction entre les relations de type famille et série est effectuée à partir des catégories grammaticales des mots.

La dernière étape du traitement correspond à l'application de trois filtres sur le graphe de relations morphologiques obtenu précédemment. Le premier filtre consiste à confronter les analogies validées à partir des formes graphiques des mots aux proportions analogiques constituées des transcriptions phonétiques des mêmes mots. Si l'analogie entre transcriptions n'est pas vérifiée, l'analogie entre formes graphiques est invalidée. Le second filtre concerne le nombre d'analogies impliquant chaque couple de mots. Si une relation entre deux mots est établie par au moins dix analogies, les deux mots sont considérés comme appartenant à la même famille. Le dernier filtre concerne la topologie du graphe. Il s'appuie sur l'idée que les séries en constituent des zones denses et que les sommets associés par une relation de type série doivent tous être associés entre eux par cette relation. Les sommets reliés seulement à une sous-partie sont exclus.

La ressource finale comporte 29 310 mots, 96 107 relations de type famille morphologique dérivationnelle et 1 160 098 relations de type série dérivationnelle. Un travail est en cours, présenté dans Hathout et Namer (2014), pour coupler cette ressource à l'analyseur du lexique morphologiquement construit du Français Dérif[22] dans le cadre d'une nouvelle ressource, Démonette. Dans ces travaux, l'analogie est considérée comme un guide, permettant de rendre compte de l'organisation du lexique.

### 1.5.3 Enrichissement de ressources lexicales

Des méthodes comparables à celle utilisée pour la constitution de Morphonette ont été utilisées par ailleurs pour l'enrichissement de ressources lexicales.

#### 1.5.3.1 Terminologie

En terminologie, citons Claveau et L'Homme (2005), qui s'appuient également sur l'hypothèse d'une corrélation entre analogie formelle et analogie sémantique.

[22] `http ://www.cnrtl.fr/outils/DeriF/`

La ressource qu'ils exploitent et enrichissent est formalisée selon les principes de la Lexicologie Explicative et Combinatoire de Mel'čuk et al. (1995), tout comme celle que nous présenterons au chapitre 2. Il s'agit d'un dictionnaire français du domaine de l'informatique, DiCoInfo[23]. Dans cette ressource, les termes entretiennent des relations sémantiques typées. Par exemple, *programme* est relié à *programmer* par un lien qui spécifie qu'il en est le résultat typique.

L'approche qu'ils adoptent est une approche de type raisonnement à partir de cas, tel qu'introduit par Kolodner (1992). Ils utilisent la ressource DiCoInfo[24], en cours de développement, pour apprendre des règles associant des variations formelles à des types de liens sémantiques. Ces règles sont apprises par abstraction à partir d'analogies formelles. Par exemple, étant donné les quatre entrées *connecteur*, *connecter*, *éditeur* et *éditer*, pour lesquelles la ressource encode les relations « *connecteur* est l'instrument typique de *connecter* » et « *éditeur* est l'instrument typique de *éditer* ». L'analogie *connecteur* : *connecter* :: *éditeur* : *éditer* aboutit à la création de la règle $\mathbf{S_{instr}}(\mathrm{m_1}) = \mathrm{m_2}$ si $\mathrm{m_1}$ -$_{\mathrm{suf}}$"eur" +$_{\mathrm{suf}}$"er" $= \mathrm{m_2}$, où $\mathbf{S_{instr}}$ représente la relation sémantique « est l'instrument typique de », -$_{\mathrm{suf}}$ la suppression d'un suffixe, +$_{\mathrm{suf}}$ son ajout et $\mathrm{m_1}$, $\mathrm{m_2}$ chacun un mot.

Parallèlement à l'apprentissage de règles, ils effectuent une extraction terminologique en corpus à l'aide du système TermoStat[25] sur un corpus de spécialité, regroupant des textes de différents sous-domaines informatiques. Ils ne conservent de cette extraction que les candidats termes simples, seuls les mots ayant un indice de spécificité supérieur à 0 étant considérés comme candidats[26].

L'ensemble des couples de candidats termes possibles est alors considéré comme un ensemble de cas inconnus, dont on chercherait à savoir s'ils correspondent à des cas déjà rencontrés auparavant. Pour se faire, chaque couple est comparé aux règles apprises lors de la première étape. Si l'alternance formelle entre les deux mots d'un couple correspond à la partie droite d'une règle, le couple est intégré à la ressource et la relation sémantique de la partie gauche de la règle est encodée.

Les résultats obtenus sont encourageants. La méthode permet de collecter 71,77% des couples de mots en relation sémantique disponibles dans le corpus de spécialité. Les couples manquants correspondent à des cas où l'un des termes n'a pas été repéré par TermoStat comme étant spécifique au domaine de l'informatique et à des cas où l'alternance graphique entre les termes était absente de la ressource initiale, comme celle qui existe entre *connecter* et *interconnexion*. Les relations induites, pour leur part, sont correctes à 65,48%. Les auteurs précisent que, dans la majorité des cas, il existe bien une relation sémantique entre les mots reliés à tort, mais qu'elle est différente de celle encodée automatiquement. Ils considèrent leur hypothèse selon laquelle « une proximité morphologique indique souvent une proximité sémantique,

[23] `http ://olst.ling.umontreal.ca/cgi-bin/dicoinfo/search.cgi`

[24] Ils testent également leur méthode sur une ressource de langue générale, mais les résultats obtenus sont moins satisfaisants.

[25] `http ://olst.ling.umontreal.ca/?page_id=91`

[26] L'indice de spécificité d'un mot est calculé par TermoStat en comparant sa fréquence dans le corpus de spécialité qui lui est fourni en entrée à un corpus, considéré comme généraliste, tiré du journal *Le Monde*.

notamment dans des domaines spécialisés » confirmée. Une telle méthode pourrait donc trouver sa place dans un contexte d'instrumentation du travail terminologique.

#### 1.5.3.2 Lexiques bilingues

Comme nous l'avons évoqué plus haut, plusieurs travaux ont été menés visant à exploiter l'analogie dans le cadre de la traduction automatique. Nous nous intéressons ici au cas particulier de Langlais et Patry (2008), qui se préoccupent, tout comme Claveau et L'Homme (2005), de l'enrichissement de ressources lexicales par exploitation d'analogies formelles. Plus précisément, ils montrent « que le raisonnement analogique offre également une réponse adéquate au problème concret de la traduction d'entrées lexicales inconnues pour différentes directions de traduction », pour peu que ces entrées lexicales soient morphologiquement liées à des entrées connues.

La méthode mise en place par Langlais et Patry s'inspire des travaux sur l'apprentissage par analogie menés par Pirrelli et Yvon (1999). Les ressources qu'ils souhaitent enrichir sont des lexiques bilingues. Chacun d'entre eux est constitué de paires de chaînes de caractères, dans lesquelles un mot d'une langue source est associé à un mot d'une langue cible. Ils fournissent en exemple le couple (*analogie*, *analogy*). Ces lexiques sont obtenus, pour chaque couple de langues[27], de manière automatique, par exploitation de bitextes[28]. Ils distinguent alors un espace d'entrée, correspondant à la langue source, d'un espace de sortie, correspondant à la langue cible. À la suite de Stroppa (2005), ils considèrent qu'« à une analogie dans l'espace d'entrée correspond une analogie dans l'espace de sortie », ce qu'ils nomment *biais analogique* et qui recoupe la notion d'homomorphisme entre structures analogiques introduite par Lepage (2003). Ils mettent alors en place une méthode en trois étapes, implémentée sous le nom d'ANALOG.

Étant donné un mot de la langue source inconnu du lexique bilingue existant, la première étape consiste à trouver, dans l'espace d'entrée, l'ensemble des équations analogiques qui mettent en jeu ce mot et dont une solution possible est un mot déjà connu. Par exemple, si les mots *insérés*, *réinsérés* et *intégrés* sont connus, mais que *réintégrés* ne l'est pas, l'équation analogique formelle *réinsérés* : *insérés* :: *réintégrés* : $x$ a pour solution possible dans l'espace d'entrée *intégrer*. Afin de réduire la complexité de ce traitement, la recherche des mots connus entrant en relation analogique avec le mot inconnu est limitée aux mots qui lui sont proches, selon la distance de Levenshtein. À cette étape, tous les triplets de mots connus entrant en relation analogique avec le mot inconnu sont conservés.

La seconde étape consiste à résoudre l'ensemble des équations analogiques formelles mettant en jeu les traductions des triplets de mots connus obtenus précédemment. Ainsi, les trois entrées (*insérés, inserted*), (*réinsérés, reinserted*), (*intégrés, incorporated*) donnent lieu à l'équation *inserted* : *reinserted* :: *incorporated* : $x$. La résolution de cette équation aboutit à plusieurs résultats, qui ne sont pas tous des mots valides de la langue cible. Il en va de même pour les autres équations obtenues

[27] Les couples de langues concernés sont français-anglais, espagnol-anglais et allemand-anglais.
[28] Un bitexte est une paire de textes tels que l'un des textes est la traduction de l'autre.

pour le mot *réintégrés*.

Lors de la troisième étape, les candidats à la traduction sont filtrés à l'aide d'un lexique monolingue à large couverture. Les formes trouvées dans ce lexique sont ensuite ordonnées selon le nombre d'équations analogiques dont elles sont une solution. Nous reprenons dans le tableau 1.3 les résultats obtenus pour le mot inconnu *réintégrés* ; **cand** indique le nombre de candidats à la traduction obtenu en sortie de la seconde étape et $\mathcal{V}$ le nombre de candidats conservés après filtrage.

| source | **cand** | $\mathcal{V}$ | (candidat, fréquence) |
|---|---|---|---|
| *réintégrés* | 2686 | 18 | (*reinstated*, 20) (*reintegrated*, 17) (*re-integrated*, 13) (*reentered*, 10) (*reincluded*, 8) (*reinvolved*, 8) (*re-incorporated*, 8) (*reinserted*, 7) (*reinstate*, 7) (*reintegrate*, 6) (*reinstating*, 4) (*accomplished*, 3) (*rebuilt*, 3) (*reinclude*, 3) (*rejoined*, 3) (*reverte*, 2) (*reintegration*, 2) (*reintegrating*, 2) |

TAB. 1.3 : Traductions de *réintégrés* proposées par ANALOG

L'évaluation d'ANALOG montre l'apport de la méthodologie proposée. Tandis que les mots inconnus ne sont généralement pas traités lors des tâches de traduction automatique, Langlais et Patry (2008) constatent qu'une traduction est proposée par leur système pour 60% des mots inconnus rencontrés dans le corpus d'évaluation de l'atelier WMT'06 (Koehn et Monz, 2006)[29], dès lors que leur lexique d'amorce est appris sur plus de 50 000 paires de phrases. De plus, pour les traductions disponibles, 81% des traductions en tête sont bonnes. 73% des mots n'ayant pas reçu de traduction satisfaisante n'en ont pas reçu du tout. Il s'agit majoritairement de noms propres, de mots d'une langue autre que la langue source et de mots composés. Ils constatent également que leur système est plus apte à proposer des traductions pour les mots peu fréquents que les approches statistiques. La méthode proposée ici semble donc, elle aussi, tout à fait apte à trouver sa place dans un contexte d'instrumentation du travail, non plus terminologique, mais de traduction.

### 1.5.4 Résolution d'analogies lexicales

Les approches que nous avons vues jusqu'à présent se concentrent toutes sur la présence d'analogies formelles dans l'organisation du lexique. Si elles parviennent également à détecter des analogies sémantiques, elles se cantonnent au cas où les mots en relations entretiennent une certaine parenté morphologique. D'autres analogies lexicales ont cependant fait l'objet de travaux de recherche. C'est le cas notamment des travaux portant sur la résolution d'analogies semblables à celles utilisées dans les tests d'aptitude scolaires aux États-Unis jusqu'au début des années 2000. Différentes approches sont proposées pour ces résolutions, qui passent par la mise au point de mesures de similarité sémantique basées sur l'exploitation de corpus et/ou de ressources lexicales.

[29]L'atelier WMT'06 mettait à disposition des bitextes d'entraînement regroupant 700 000 paires de phrases extraites de textes parlementaires européens, ainsi que deux corpus d'évaluation, l'un constitué de 2 000 phrases du même domaine et 1 064 phrases d'autres domaines.

#### 1.5.4.1 Test d'aptitude scolaire

Le test d'aptitude scolaire, ou SAT Reasoning Test, est un examen national américain sous forme de questionnaire à choix multiples, utilisé pour l'admission à l'université. Jusqu'en 2005, ces tests comportaient des analogies lexicales[30]. Une telle analogie comprend un couple de mots source et un ensemble de couples de mots cibles, généralement cinq selon Veale (2004). L'exercice consiste à identifier, parmi les cibles proposées, le couple de mots qui entretiennent la ou les mêmes relations que les mots du couple source. En voici un exemple, adapté de Veale (2004) :

| | Osterich is to Bird as : |
|---|---|
| a) | Cub is to Bear |
| b) | *Lion is to Cat* |
| c) | Ewe is to Sheep |
| d) | Turkey is to Chicken |
| e) | Jeep is to Truck |

La bonne réponse, en italique, correspond à l'analogie *Osterich* : *Bird* :: *Lion* : *Cat*. De manière équivalente, en français, nous pouvons dire[31] que « Autruche est à Oiseau ce que Lion est à Félin ». Turney et Littman (2005) estiment que les étudiants soumis au test parviennent à reconstituer, en moyenne, 57% des analogies ainsi présentées.

#### 1.5.4.2 Mesure de similarité basée sur corpus

Turney et Littman (2005) et Turney (2006) proposent des méthodes de calcul de similarité de relations sémantiques entre couples de mots, qu'ils testent sur la résolution d'analogies lexicales. Nous décrivons ici la méthode datant de 2006, présentée comme une amélioration de celle de 2005 et qui permet d'obtenir un taux de résolution proche de celui des étudiants.

La méthode présentée dans Turney (2006), baptisée LRA pour *Latent Relational Analysis*, se décompose en sept étapes. Elle se base sur le principe, découlant des travaux de Gentner (1983) et Medin et al. (1990), que pour toute analogie $A : B :: C : D$, il existe un fort degré de similarité de Relations entre les couples (A,B) et (C,D) et qu'il est possible de distinguer des analogies proches, telles que *mason* : *stone* :: *carpenter* : *wood*, d'analogies lointaines, telles que *traffic* : *street* :: *water* : *riverbed*. Les premières correspondent au cas où il existe un fort degré de similarité d'Attributs entre A et C d'une part et B et D d'autre part. LRA prend en entrée un ensemble de couples de mots et fournit en sortie un degré de similarité de Relations entre couples. Dans les cas d'un test SAT, l'ensemble est constitué du couple source et des couples cibles proposés, en sortie, seules les similarités entre le couple source et chacun des couples cibles sont considérées.

[30] Turney (2006) nous apprend que les analogies ont été supprimées des tests car elles étaient considérées comme discriminantes à l'égard des minorités.

[31] Il faut noter que CAT, en anglais, a un sens équivalent à *félin*.

Pour Turney, deux mots qui entretiennent un fort degré de similarité d'Attributs sont synonymes. Les Relations entre les mots peuvent, elles, être observées en corpus. Afin de maximiser cette observation, il s'intéresse non seulement aux couples fournis en entrée, mais à un ensemble de couples *alternatifs* entrant en relation d'analogie avec ceux-ci et mettant en jeu des synonymes des mots qui les composent. Pour chaque couple de mots (A,B), la première étape de traitement consiste à interroger un thésaurus afin de constituer des couples (A',B) et (A,B'), dans lesquels A' et B' sont des synonymes de A et B. Il considère que ces couples entrent en relation analogique proche avec le couple initial de la manière suivante : $A : B :: A' : B$ ou $A : B :: A : B'$.

Turney se situe dans un cadre où un mot correspond à une forme et où aucune distinction n'est faite, a priori, entre les différents sens associés à cette forme. La collecte de synonymes effectuée lors de la première étape peut donc aboutir à de mauvaises analogies. Imaginons un couple fourni en entrée composé de *vache* et *veau*. L'ambiguïté lexicale du mot *vache* — qui dénote notamment un animal bovin dans son sens premier et un policier dans un sens métaphorique — peut amener à considérer l'analogie *vache* : *veau* :: *policier* : *veau*. Non seulement cette analogie n'est pas proche, mais elle est incorrecte. Afin de limiter les risques de telles mauvaises alternatives, la seconde étape du traitement consiste à chercher en corpus le nombre d'expressions courtes débutant par l'un des membres d'un couple alternatif et terminant par l'autre. Seuls les couples de mots cooccurrant fréquemment sont conservés. Dans les expériences présentées, la taille des expressions courtes est établie à cinq mots maximum et seuls les trois couples les plus fréquents sont conservés.

La troisième étape consiste également à chercher en corpus des expressions débutant par l'un des membres d'un couple de mots et terminant par l'autre. Cette opération est effectuée pour le couple initial comme pour les couples alternatifs conservés. Les expressions sont ici également limitées en taille et doivent, de plus, comporter au minimum un mot entre les mots du couple qu'elles concernent. Par exemple, pour le couple (*vache*,*veau*), une recherche dans le corpus frWaC[32] fourni, entre autres, les expressions *vache et son veau*, *vache a fait son veau*, *veau à la vache* et *veau de sa vache*. L'hypothèse de Turney est que de telles expressions caractérisent la relation entre couples de mots.

La quatrième étape consiste à transformer l'ensemble des expressions collectées en motifs, en remplaçant un, tous ou aucun des mots intermédiaires par des jokers, autant de fois que possible. Ainsi, l'expression « veau de sa vache » aboutirait aux motifs « *de sa* », « *de ** », « ** sa* » et « ** ** ». LRA calcule le nombre de couples de mots, initiaux ou alternatifs, qui cooccurrent à l'aide de chacun des motifs. Seuls les motifs les plus fréquents sont conservés.

Une matrice de fréquences est alors construite, rendant compte du nombre de fois où chaque couple de mots apparaît en corpus associé à chaque motif. Différentes transformations sont appliquées à cette matrice, ayant pour but de donner davantage de poids aux motifs pour lesquels la répartition entre couples est la moins homogène. Une version lissée et compressée de la matrice est ensuite obtenue à l'aide d'une dé-

[32] `http ://wacky.sslmit.unibo.it/doku.php?id=corpora`.

composition en valeurs singulières. Ces opérations sont réalisées lors des étapes cinq et six. La mesure de similarité entre les couples de mots initiaux est effectuée sur la matrice optimisée.

Soient (A,B) et (C,D) deux couples fournis en entrée. À cette étape du traitement, LRA dispose d'un certain nombre de couples correspondant à (A,B) — dans le cas de l'expérience détaillée un couple initial plus les trois couples alternatifs dont les membres cooccurrent le plus fréquemment en corpus — et d'autant correspondant à (C,D). Chacun de ces couples est représenté par une ligne de la matrice lissée, équivalant à un vecteur. La similarité entre les couples représentant (A,B) et ceux représentant (C,D) est alors mesurée à l'aide d'un cosinus de vecteurs. La valeur obtenue par les couples initiaux sert de valeur seuil. Les valeurs obtenues à partir de couples dont l'un au moins n'est pas un couple initial ne sont conservées que si elles lui sont supérieures. En effet, Turney estime que si l'analogie formée par des couples alternatifs a un cosinus supérieur à celui des couples initiaux, c'est qu'il a trouvé un meilleur moyen d'exprimer l'analogie, sans en modifier le sens. En revanche, si ce cosinus est inférieur, il considère que le sens a été modifié par un mauvais choix de synonymes à la première étape et que l'analogie n'est pas valide. LRA calcule alors la moyenne des cosinus conservés. La valeur de cette moyenne est considérée comme étant la similarité de relations existant entre (A,B) et (C,D).

Appliquée à 374 questions de tests SAT, la méthode LRA permet d'obtenir 210 réponses correctes. Turney mesure un taux de précision de 56,8%, un rappel de 56,1% et une F-mesure de 56,5%, ce qui est équivalant à la performance humaine estimée à 57%.

Turney (2008b) utilise une variante de LRA, qui fait notamment intervenir un algorithme d'apprentissage et l'idée selon laquelle l'identification d'analogies peut être vue comme une tâche de classification de relations sémantiques entre mots. De ce point de vue, l'analogie *mason* : *stone* :: *carpenter* : *wood* correspond à l'alternance d'étiquettes classifiantes **artisan** : **material** dans chacun des couples (*mason, stone*) et (*carpenter, wood*). Cette approche aboutit à de moins bons résultats que celle de 2006 sur la résolution des questions du test SAT, avec 52,1% de réponses correctes, mais lui permet de proposer un traitement uniforme pour un ensemble de phénomènes sémantiques. La synonymie, l'antonymie et l'association d'idées sont alors considérées comme un ensemble de cas particuliers d'analogie. Turney (2012) poursuit cette idée d'un traitement uniforme de différents phénomènes linguistiques en s'intéressant à la paraphrase de deux mots par un seul — comme dans le cas de « *dog house* » paraphrasé par « *kennel* » — conjointement à la reconnaissance de relations sémantiques. Il fait alors évoluer sa méthode pour distinguer les patrons nominaux des patrons verbaux. Il considère que les premiers caractérisent le domaine d'un mot — comme le domaine de la charpenterie pour *carpenter* — tandis que les seconds caractérisent son rôle dans le domaine — comme le rôle artisan pour ce même mot. Turney (2013) utilise une variante de cette méthode, qu'il optimise en la couplant à un algorithme d'apprentissage. Dans ce cadre, il exploite une partie des analogies équivalentes de Lepage (2003) pour accroître la quantité de données d'apprentissage. Si l'analogie *word* : *language* : *note* : *music* est connue pour être vérifiée, *language* : *word* :: *music* : *note*, *note* : *music* :: *word* : *language*

et *music* : *note* :: *language* : *word* sont également considérées comme des données valides. Selon le même principe, *word* : *music* :: *note* : *language* est utilisée comme donnée invalide. Cette méthode, baptisée *SuperSim*, obtient des résultats sensiblement équivalents à ceux de LRA sur le jeu de questions SAT, avec 54,8% de réponses correctes. Son temps d'exécution, en revanche, n'est que de quelques minutes, contre neuf jours pour LRA.

Turney (2008a), que nous avons déjà évoqué précédemment, utilise également une variante de LRA pour la résolution d'anaphores scientifiques et de métaphores courantes, vue comme des systèmes complexes de proportions analogiques.

L'ensemble de ces travaux met l'accent sur la mise au point d'une mesure de similarité sémantique et l'identification des relations en œuvre dans cette similarité. Comme le détaille Turney (2006), de tels dispositifs trouvent de nombreuses applications en traitement automatique des langues, en dehors du champ de l'analogie. Ils pourront notamment être utilisés pour la désambiguïsation lexicale, l'extraction d'information et l'identification de rôles sémantiques.

#### 1.5.4.3 Mesure d'analogicité basée sur WordNet

Veale (2004) aborde la question des analogies du test SAT sous un angle différent. Il s'agirait d'analogies conceptuelles, qu'il serait possible de résoudre à l'aide d'une base de connaissances. Il parle alors de *termes* ou de *concepts* mis en relation analogique et propose d'exploiter la ressource WordNet (Fellbaum, 1998) à cette fin, ou plus précisément la taxinomie de noms qu'elle comporte. Il baptise sa méthode KNOW-BEST, pour *KNOWledge-Based Entrainment and Scholastic Testing*. Son objectif est double : mesurer la capacité de WordNet à servir dans une telle tâche et établir les bases permettant de construire un logiciel autonome capable d'inventer, formuler et évaluer ses propres tests, tout autant que de les résoudre.

Reprenons l'exemple que nous avons introduit en 1.5.4.1. Au couple source (Osterich, Bird) était associé cinq propositions de couples : (Cub, Bear), (Lion, Cat), (Ewe, Sheep), (Turkey, Chicken) et (Jeep, Truck)[33]. Pour distinguer la bonne solution parmi ces couples, il n'est pas possible de se contenter de comparer directement les relations qui les relient dans WordNet. En effet, la relation d'hyperonymie relie tout aussi bien (Lion, Cat) que (Ewe, Sheep). Une relation plus fine, de type « est le plus grand des » serait nécessaire, mais n'est pas disponible.

Veale (2004) utilise toutefois les relations brutes de WordNet pour filtrer les propositions. Dans un premier temps, il rapporte, quand cela est rendu possible par la présence de liens morpho-sémantiques, tous les termes qui ne sont pas nominaux à des noms. Par exemple, *serene* devient *serenity*. Il justifie ce choix par le fait que la taxinomie nominale est la partie la plus riche de la ressource et qu'elle comporte des relations de subsomption. Chaque couple de termes nominaux subit ensuite une analyse pour déterminer si ses membres sont reliés dans WordNet. Les couples de termes qui ne le sont pas sont éliminés. De plus, si le couple source implique une

[33]Une traduction possible de ces couples est la suivante : (Autruche, Oiseau), (Ourson, Ours), (Lion, Félin), (Brebis, Mouton), (Dinde, Poulet), (Jeep, Camion).

relation de subsomption, seuls les couples de candidats qui en impliquent également une sont conservés.

Les couples restants sont associés, un à un, au couple source. Ils forment ainsi des analogies potentielles, telles que Osterich : Bird :: Ewe : Sheep, qui sont soumises à une mesure de *similarité analogique*. Le couple cible dont l'association obtient le meilleur score est considéré comme étant la solution.

La mesure de similarité analogique est effectuée en plusieurs étapes. En premier lieu, chaque couple de termes est considéré comme un couple de concepts et se voit attribuer une mesure de *similarité taxinomique*. Cette mesure est basée sur la profondeur de chaque concept dans la taxinomie, la profondeur de leur plus proche parent commun et le nombre de termes adjectivaux qui sont partagés par leurs gloses. La notion de plus proche parent commun est affinée. Tandis qu'on considère généralement qu'il s'agit de l'hyperonyme le plus spécifique des deux concepts, Veale considère que deux hyperonymes qui ont une lexicalisation similaire sont unifiables. Ainsi, `{reproductive_structure}` et `{reproductive_cell}` sont considérés comme unifiables et c'est leur profondeur qui est prise en compte pour le parent commun d'un couple de concepts tel que (Seed, Egg), qui sont donc considérés comme davantage similaires que dans le cas où un parent commun serait à chercher plus loin. La similarité de l'analogie formée de deux couples est calculée en effectuant la somme de leurs similarités taxinomiques. Cette mesure est ensuite pondérée à l'aide d'un critère basé sur l'*hypothèse d'invariance*.

À la suite de Lakoff (1987), Veale postule l'existence d'une structuration des domaines en « schémas-images », correspondant à un certain nombre de catégories générales disponibles dans une ontologie, comme INSTRUMENT ou SUBSTANCE. Il considère que, dans le cas de l'analogie, l'hypothèse d'invariance implique que si l'un des membres d'un couple appartient à l'une de ces catégories, le second membre y appartient également. Il choisit sept catégories générales[34] et comptabilise le nombre de variations de catégories en œuvre dans chaque analogie. La similarité d'analogicité est diminuée d'autant que ce nombre est grand.

Veale (2004) teste les performances de sa méthode sur le même corpus de 374 analogies que Turney (2006). Il obtient des réponses pour l'ensemble des cas, mais seulement 42% sont correctes. Ces résultats sont supérieurs à ceux qu'il observe dans le cas d'une comparaison littérale des couples, mais inférieurs à ceux de Turney (2006). Il suggère d'exploiter sa méthode en complément d'une méthode d'extraction d'informations. Un tel couplage permettrait, selon lui, de s'attaquer à la question de l'identification des relations en jeu dans chacun des couples entrant en relation analogique.

Veale (2006) approfondit la question de l'utilisation de WordNet pour générer des analogies. Il développe l'idée qu'il est possible d'y identifier des dimensions conceptuelles, lexicalisées ou *ad hoc*, à partir desquelles la structure analogique peut-être imposée à son contenu lexical. Une analogie telle que Poseidon:Sea::Apollo:Sun se-

[34] INSTRUMENT, COLLECTION/GROUP, LOCATION, ANIMAL/PERSON, ROLE, SUBSTANCE et CONTAINER.

rait bâtie sur une dimension ad hoc de type « choses qui peuvent être personnifiée comme dieux » à laquelle « Sea » et « Sun » appartiendraient, tandis que l'analogie Astronaut:Spacecraft::Airman:Aircraft serait basée sur une dimension lexicalisée « Conveyance », à laquelle « Spacecraft » et « Aircraft » sont d'ores et déjà associés dans la ressource.

Parallèlement à ces travaux, Veale (2003) s'intéresse à la constitution de thésaurus analogique, rendant davantage compte des différentes dimensions structurant le lexique et des rapports en jeu dans les analogies, qu'il qualifie de *pivots*. Il utilise WordNet comme base de ce thésaurus et l'enrichit de nouvelles catégories. Ainsi, tandis qu'il dispose au départ d'une hiérarchie des divinités organisée par religion (divinités grecques, divinités hindoues, etc.), il ajoute une seconde hiérarchie rendant compte des symboles partagés par ces divinités à travers les religions. « Zeus » et « Veruna » sont ainsi associés à la nouvelle catégorie `{Supreme_deity}`. Le rapport mis en évidence par l'analogie « Zeus est à la mythologie grecque ce que Veruna est à la religion hindoue », « son dieu suprême », est alors directement accessible. Cette idée de ressource structurée par l'analogie est étudiée sous un angle différent par Veale et Li (2014). Ils discutent la pertinence d'un enrichissement automatique de bases de connaissances en imposant que chaque nouvelle entrée entre en relation analogique conceptuelle avec les entrées existantes. Ils défendent l'idée qu'une telle approche permet de s'assurer de la qualité de la ressource, dont la structure devrait alors être équilibrée, cohérente et riche en isomorphismes. Ils montrent également comment leur méthodologie aboutit à une acquisition rapide et simplifiée de nouveaux concepts à partir de corpus.

Ces travaux posent également la question d'une mesure de similarité de relations sémantiques, mais sous un angle différent. C'est ici la ressource ontologique qui est au cœur des préoccupations. Pour être de qualité, cette ressource doit rendre compte correctement de la structuration des concepts. Elle doit permettre d'accéder à un haut niveau de généralisation à partir de la représentation qu'elle offre de chaque domaine de spécialisation et encoder l'ensemble des rapports mis en œuvre lorsqu'une analogie est effectuée entre domaines.

# Conclusion

Nous avons vu, dans la section 1.5, plusieurs exemples d'informatisation du raisonnement analogique en traitement automatique des langues. Nous nous sommes focalisée sur un ensemble d'expérimentations rendant compte du lien entre analogie et organisation du lexique. Les méthodes présentées qui exploitent l'analogie formelle mettent en avant la possibilité d'extraire des observables structurés à partir de corpus. Elles ne permettent pas de couvrir l'ensemble des cas rencontrés, car elles se limitent aux unités morphologiquement liées, mais confortent néanmoins l'hypothèse de Lepage (2003) d'un homomorphisme entre niveaux de représentations linguistiques et entre langues. Le bénéfice de leur exploitation dans le cadre d'une instrumentation du travail sur les ressources lexicales a également été esquissé. D'autres méthodes présentées cherchent à reproduire directement l'analogie sémantique, sans passer par une étape graphique. Ces méthodes se heurtent à la difficulté de formaliser les rapports en jeu dans de telles analogies et de mesurer

leurs similarités. Elles proposent des solutions basées sur l'exploitation de corpus ou de ressources ontologiques et obtiennent des résultats satisfaisants. Dans l'ensemble des cas de figure rencontrés, l'analogie est perçue comme un guide pour la structuration du lexique, permettant d'en assurer une représentation de qualité et d'appréhender la question de l'intégration de nouvelles données. D'autres travaux s'intéressent à l'exploitation de ressources lexicales et à l'analogie. Nous pensons notamment aux travaux de Gaume (2003), Duvignau et Gaume (2004) et Desalle (2012), en psycholinguistique, qui n'ont pas trouvé leur place dans ce chapitre, mais qui ont également joué un rôle dans notre familiarisation avec la notion.

La question qui nous préoccupe dans le cadre de notre travail de thèse est légèrement différente des tâches que nous avons présentées. Nous pensons que les qualités de structuration de représentation des connaissances relevées par Veale doivent également s'appliquer aux ressources lexicographiques. Comme nous l'avons dit précédemment, l'intuition d'un locuteur français d'une analogie entre mots doit être affinée par les descriptions lexicographiques mises à sa disposition. Une ressource qui rend compte explicitement des rapports en jeu dans cette analogie et de la structuration des unités lexicales qui la sous-tend pourra être exploitée dans de nombreuses tâches de traitement automatique des langues. Elle trouvera également sa place en pédagogie, permettant notamment de travailler sur l'analogie et l'abstraction dans un cadre lexical. Une ressource lexicographique du français d'envergure est en cours d'élaboration au laboratoire d'Analyse et de Traitement Informatique de la Langue Française, ATILF. Construite selon les principes de la Lexicologie Explicative et Combinatoire de Mel'čuk et al. (1995), cette ressource encode de nombreux liens sémantiques entre unités lexicales. Nous pensons que l'exploration de cette ressource par raisonnement analogique permettra d'acquérir automatiquement des connaissances sur l'organisation du lexique. Ces connaissances pourront être exploitées pour identifier des phénomènes linguistiques et instrumenter l'activité lexicographique.

Comme nous l'avons vu au fil de ce chapitre, la tâche d'exploration automatique par raisonnement analogique nécessite une réflexion sur la nature des objets mis en relation et la mise au point de mesures de similarités d'Attributs et de Relations entre objets. Ces questions nous occuperont tout au long de notre réflexion. Les questions d'optimisation de traitements, mises en évidence par les méthodes que nous avons présentées, ne seront pas entièrement résolues ici. Le travail que nous présentons est avant tout exploratoire et son informatisation n'est pour l'heure qu'à l'état de prototype. Nous entendons ici présenter une première étape, permettant de poser les bases nécessaires à la réalisation de travaux futurs, tels que l'enrichissement semi-automatique de la ressource.

# Chapitre 2

# Réseau Lexical du Français (RL-fr)

## Sommaire

# Introduction

La mise en œuvre d'une tâche d'informatisation du raisonnement analogique nécessite une réflexion préalable sur la nature des objets que l'on souhaite comparer et des rapports qu'ils entretiennent. Ce second chapitre présente la ressource lexicale que nous souhaitons explorer à l'aide du raisonnement analogique.

Le Réseau Lexical du Français, désormais RL-fr, s'inscrit dans la lignée de l'approche lexicographique des dictionnaires virtuels et des travaux en Lexicologie Explicative et Combinatoire. Sa structuration suit le modèle de système lexical introduit par Polguère (2009, 2014a,b). Il s'agit d'un graphe orienté, encapsulé dans une base de données contenant des entités et des relations de natures variées.

Après une présentation générale des systèmes lexicaux, ce chapitre détaille les différents éléments du RL-fr. Les propriétés formelles de ce graphe lexical sont ensuite présentées. Lors de cette analyse topologique, la question de l'appartenance du RL-fr à la catégorie des graphes petit monde est abordée.

## 2.1 Systèmes lexicaux

Comme nous l'avons évoqué précédemment, une ressource lexicographique du français est en cours d'élaboration au laboratoire ATILF, le RL-fr. Cette tâche lexicographique a débuté en 2011. Elle s'inscrit dans le courant d'une lexicographie contemporaine, telle que celle décrite par Atkins (1996), Selva et al. (2003) et Spohr (2012). Tout en s'appuyant sur les points forts de la lexicographie traditionnelle, celle-ci profite des avancées de l'informatique et des théories linguistiques pour rompre avec le modèle dictionnairique classique. L'objet réalisé n'est plus une succession de descriptions textuelles organisée par ordre alphabétique, mais un graphe lexical, encapsulé dans une base de données lexicales structurée. Afin d'assurer l'homogénéité et l'intégrité de la ressource, le recours à l'écriture de chaînes de caractères par les lexicographes est limité autant que possible. Leur travail s'effectue à l'aide d'un éditeur, Dicet (Gader et al., 2012). Chaque élément créé dans la base de données, tel qu'une caractéristique grammaticale ou une unité lexicale, est associé à un identifiant unique. C'est cet identifiant qui est mis en relation avec les autres éléments de la base. Pour les lexicographes, la tâche de description consiste à sélectionner les éléments souhaités dans l'éditeur, à tisser des liens entre les unités lexicales et leurs éléments de description. Un ensemble de dictionnaires répondant à des besoins spécifiques, tout autant que de lexiques pour le traitement automatique des langues, peut être généré à partir de cette ressource. Il s'agit en cela d'une lexicographie des dictionnaires virtuels.

### 2.1.1 Réseaux lexicaux

En tant que graphes, les systèmes lexicaux sont proches des réseaux que sont les différents WordNet (Fellbaum, 1998) et BabelNet (Navigli et Ponzetto, 2010). Le lexique y est représenté par un ensemble de sommets reliés entre eux par un ensemble de relations typées. Ils s'en distinguent cependant sur plusieurs points. Premièrement, ils ne se cantonnent pas aux parties du discours majeures des noms,

verbes, adjectifs et adverbes, qu'il s'agirait de traiter indépendamment les unes des autres. L'ensemble des unités lexicales, y compris les mots grammaticaux, y trouve sa place et y est décrit selon le même formalisme. Bien qu'elle soit représentée, la synonymie n'occupe pas une place prépondérante dans l'organisation des systèmes lexicaux. De plus, les relations encodées n'y sont pas majoritairement hiérarchiques. Enfin, la visée des systèmes lexicaux est exclusivement lexicale. Ils ne cherchent aucunement à rendre compte de l'organisation de concepts lexicalisés, ni d'aucune connaissance encyclopédique sur le monde.

Soulignons également qu'aucun système lexical multilingue n'est envisagé. Si la création de dictionnaires ou de lexiques multilingues à partir de tels systèmes n'est pas exclue, elle s'envisage à partir de versions monolingues, à la manière de Spohr et Heid (2006). À l'heure actuelle, le RL-fr est le système lexical le plus avancé en terme de finesse de description. Au cours des trois dernières années, trois autres réseaux lexicaux ont été initiés. Le RL-ko, pour le coréen, est constitué dans le cadre d'un doctorat, depuis 2011, dans une perspective de lexicologie comparative. Il se concentre autour des noms d'éléments du corps (Kim, 2013). Le RL-es, pour l'espagnol, fait l'objet d'un second doctorat, débuté en 2010 (González Orellana, 2012). Ces réseaux représentent des échantillons peu étendus, mais qui ont permis une réflexion poussée sur l'adaptation de l'éditeur et du formalisme lexicographiques à d'autres langues. Le RL-en, pour la langue anglaise, a été ébauché selon une toute autre méthodologie. Il a été généré automatiquement à partir du WordNet de Princeton (Gader et al., 2014b). Il dispose donc d'une large couverture, mais nécessite un travail de description lexicographique minutieux pour être rendu conforme au modèle. Le tableau 2.1 présente le nombre de sommets et d'arcs que comptent les systèmes lexicaux, en date du 25 août 2014. Un RL-ar, pour l'arabe et un RL-ru, pour le russe, ont également été amorcés récemment.

| | RL-fr | RL-ko | RL-es | RL-en |
|---|---|---|---|---|
| Sommets | 23 661 | 547 | 1 011 | 206 976 |
| Arcs | 53 414 | 470 | 1 072 | 946 208 |

TAB. 2.1 : Étendue des différents systèmes lexicaux

Un autre réseau lexical, JeuxDeMots (Lafourcade et Joubert, 2008, 2010), peut également être comparé aux systèmes lexicaux. Tout comme ces derniers, JeuxDeMots propose une modélisation du lexique dans laquelle l'ensemble des unités lexicales trouvent leur place, y compris les mots grammaticaux. De plus, son organisation n'est pas hiérarchique et les relations dont il rend compte sont de natures variées. Il s'en distingue cependant sur plusieurs points. Avant tout, cette ressource n'est pas le fruit d'un travail d'experts. Elle est constituée de manière collaborative, à l'aide d'un jeu en ligne auquel tout le monde peut prendre part. Les relations lexicales qu'elle réunit ne sont pas le fruit d'une description lexicographique de chaque unité lexicale, mais d'associations effectuées par des joueurs. La validité de ces relations est assurée par la redondance de mêmes propositions par différents joueurs. Un poids est par ailleurs attribué à chacune d'entre elles, en fonction de cette redondance.

### 2.1.2 Graphes d'unités lexicales

Selon Polguère (2014b), les systèmes lexicaux se distinguent des autres modèles lexicaux par la concomitance de quatre caractéristiques principales. Les deux premières concernent les éléments constitutifs du graphe, ses arcs et ses sommets. Comme nous le détaillons dans la section 2.2, les relations encodées dans les systèmes lexicaux sont orientées, c'est pourquoi nous parlons d'arcs plutôt que d'arêtes. La majorité d'entre elles sont des liens sémantico-syntaxiques de fonctions lexicales Sens-Texte (Mel'čuk, 1996). Elles s'établissent entre des unités lexicales.

Conformément au cadre de la Lexicologie Explicative et Combinatoire, une unité lexicale s'entend ici comme une *lexie*, ayant un sens, un signifiant linguistique et un ensemble de traits de combinatoire (Mel'čuk et al., 1995, p.16). L'approche classique d'une entrée de dictionnaire regroupant différents sens d'une même unité est ici abandonnée au profit d'une approche consacrant une entrée indépendante à chaque sens. Le regroupement des lexies partageant le même signifiant et liées sémantiquement reste toutefois accessible par le biais de la notion de *vocable*. Un vocable est dit *monosémique* s'il ne comporte qu'une seule lexie, *polysémique* autrement. Les lexies partageant le même signifiant, mais aucun lien sémantique sont traitées comme des *homonymes* et réparties dans des vocables distincts. Cette granularité distingue les systèmes lexicaux des graphes issus de dictionnaires papier exploités par Gaume (2004), Gaillard et al. (2011) ou encore Loiseau et al. (2011), dont les sommets sont des formes phoniques/graphiques. Il est important de noter que les vocables ne sont pas des éléments du graphe lexical. Ils sont accessibles à partir de la structure interne de chaque sommet. C'est là la seconde caractéristique des systèmes lexicaux. Leurs sommets lexicaux ne sont pas atomiques, mais contiennent une description lexicographique complexe.

### 2.1.3 Organisation du lexique

La troisième caractéristique des systèmes lexicaux concerne l'organisation du lexique qu'elle reflète. Contrairement aux modèles ontologiques, cette organisation se base tout autant sur des relations syntagmatiques que paradigmatiques. Le développement de la ressource ne s'effectue pas autour de classes ou de sous-classes préétablies, relevant d'une partie du discours ou d'un domaine particulier. Polguère et Sikora (2013) distinguent trois étapes dans le développement de la nomenclature d'un système lexical. Une *nomenclature d'amorçage* est avant tout mise au point. Elle est constituée de vocables, considérés comme le noyau lexical minimal de la langue à décrire. Les lexicographes identifient grossièrement l'*unité lexicale de base* de chacun de ces vocables, la lexie dont ils considèrent qu'elle en est le « noyau sémantique ». La structure polysémique du vocable se développera, plus tard, autour de cette unité. Ils établissent alors une *nomenclature directement induite* en tissant des liens entre chacune de ces lexies et l'ensemble de leurs dérivés sémantiques proches, tels que synonymes exacts, antonymes exacts, nominalisations et verbalisations. Un travail lexicographique de précision débute alors, au cours duquel la description de la combinatoire lexicale et du paradigme de chaque lexie amène à en introduire de nouvelles, formant une *nomenclature indirectement induite*.

Cette apparente absence de structure classifiante est cependant contrebalancée

par les propriétés formelles des systèmes lexicaux. En effet, celles-ci doivent être conformes aux propriétés des graphes « petit monde ». Comme nous le détaillons dans la section 2.3.2, de tels graphes rendent compte d'une organisation en espaces structurés. Alors que leurs sommets sont relativement peu interconnectés, ils s'organisent en zones denses. Ces zones sont reliées par le biais de sommets carrefours, permettant un accès rapide à l'ensemble des sommets depuis n'importe lequel d'entre eux. Dans le cas de graphes lexicaux, ils mettent donc en avant l'existence d'espaces sémantiques, tels que celui formé par les lexies RESSEMBLANCE **1**, SIMILITUDE, ANALOGIE **b**, SIMILARITÉ, DIFFÉRENCE, DISSIMILARITÉ, RESSEMBLANCE **2**, POINT COMMUN, FRAPPANT, CONTRASTE, SIMILAIRE, PRÉSENTER **II.2b**, DISSEMBLANCE **1** et ÉVIDENT, ainsi que de lexies carrefours, telles que FAIRE **II.1** [*Il fait du ping-pong.*], jouant un rôle central dans l'organisation du lexique.

De nombreux travaux se sont intéressés à l'exploration de graphes lexicaux à l'aide de méthodes mathématiques pour en extraire une connaissance sémantique sous-jacente. Ainsi, Ploux et Victorri (1998), Gaume (2004) et Lafourcade et Joubert (2008) proposent un traitement de la polysémie par délimitations d'espaces sémantiques et Desalle et al. (2014b) une méthode de désambiguïsation lexicale. Dans un autre contexte, Duvignau et Gaume (2004), Desalle (2012) et Desalle et al. (2014a) montrent l'apport de telles explorations en psycholinguistique.

### 2.1.4 Indice de confiance

La quatrième caractéristique des systèmes lexicaux est la présence d'une mesure de confiance associée à chaque information lexicale encodée, qu'il s'agisse des sommets et des arcs du graphe ou des éléments de descriptions lexicographiques associés aux sommets. Cette mesure prend la forme d'un pourcentage. Les lexicographes disposent du haut de l'échelle, entre 60% et 100%, pour estimer la fiabilité des informations qu'ils encodent. Dans le cadre de leur travail, cette fonctionnalité leur permet de mettre de côté les cas problématiques dont ils souhaitent discuter collectivement ou qui nécessitent la création de nouveaux éléments de description, telle qu'une caractéristique grammaticale absente de la base de données. Le bas de l'échelle est consacré aux enrichissements automatiques. Ainsi, toutes les informations initiales du RL-en ont un indice de confiance de 50%, puisqu'elles ont été générées à partir du WordNet de Princeton. Seul un travail lexicographique manuel permettra de valider son contenu et de faire évoluer ces indices.

## 2.2 Éléments d'informations lexicales du RL-fr

Nous venons de présenter les caractéristiques générales des systèmes lexicaux. Comme nous l'avons vu, il s'agit de graphes de lexies, reliées par des relations de natures variées et contenant chacune une description lexicographique complexe. Ces graphes ont les propriétés formelles de graphes petit monde et chaque information qu'ils encodent est associée à un indice de confiance. Nous allons à présent nous concentrer sur le cas particulier du système lexical du français, le RL-fr.

Depuis juin 2011, une dizaine de lexicographes travaillent à son élaboration. En tant que premier système lexical, son développement s'accompagne de la conception

et de la réalisation de l'éditeur lexicographique Dicet, qui intégre progressivement l'ensemble des fonctionnalités nécessaires. Nous nous en tiendrons ici à la présentation des éléments d'informations lexicales d'ores et déjà disponibles. Ainsi, certains aspects, tels que la paraphrase définitionnelle, seront simplement évoqués. Jusqu'à présent, seules quelques informations morphologiques sont déduites automatiquement[1], qui font l'objet d'une révision hebdomadaire. La quasi-totalité des informations lexicales contenues dans la ressource dispose donc d'indices de confiance situés dans le haut de l'échelle, entre 60% et 100%.

La présente section introduit, de manière plus fine que précédemment, la nature des sommets et des arcs encodés dans le RL-fr. Elle offre ensuite un panorama des différents éléments constitutifs des descriptions lexicographiques encapsulées dans chaque sommet.

### 2.2.1 Sommets lexicaux

Comme nous l'avons dit précédemment, les sommets du RL-fr sont, dans leur grande majorité, des unités lexicales spécifiques, des lexies, ayant un sens, un signifiant linguistique et un ensemble de traits de combinatoire. Chaque lexie regroupe un ensemble de mots-formes ou syntagmes qui correspondent à sa flexion.

Il revient au lexicographe de délimiter avec rigueur les différentes lexies correspondant aux mêmes signifiants. Il effectue ce travail selon les principes de la Lexicologie Explicative et Combinatoire (Mel'čuk et al., 1995 ; Polguère, 2011b). À partir de son intuition de l'existence d'une lexie, il rassemble des exemples dans lesquels elle apparaît. Il dispose alors de cinq critères pour guider son intuition et lui permettre de déterminer si ces exemples relèvent bien d'une seule et même lexie. Ces critères doivent être pris en considération de manière systématique et concomitante.

Sans entrer dans les détails de ces critères, intéressons-nous au premier d'entre eux. Il s'agit du *critère d'interprétation multiple*. S'il est possible de construire une phrase dans laquelle la lexie hypothétique étudiée fait l'objet d'une interprétation multiple, alors son occurrence est ambigüe et correspond à plusieurs lexies. Mel'čuk et al. (1995) fournissent l'exemple de « *Jean a peint le plafond* ». Il est possible d'interpréter *a peint* dans cette phrase de deux façons différentes et d'en conclure soit que Jean a produit une œuvre d'art, soit qu'il a fait quelques travaux de ravalement. Ces deux interprétations sont toutefois liées sémantiquement. Le lexicographe doit donc s'intéresser à au moins deux lexies pour un même vocable PEINDRE.

Nous avons jusqu'à présent considéré l'ensemble des sommets du RL-fr comme un tout homogène, désigné sous le terme de lexie. Il est cependant possible de distinguer plusieurs types de sommets, dont la description lexicographique ne suit pas exactement le même modèle. Ces types sont au nombre de trois.

Les unités monolexématiques sont appelées *lexèmes* et sont toutes des unités lexicales. Selon Mel'čuk et al. (1995), « un **lexème** est un mot pris dans une seule

[1]Les informations morphologiques dont il est question ici sont des informations flexionnelles. Nous présentons cet aspect dans la section 2.2.4.

acception bien déterminée et munie de tous les renseignements qui spécifient totalement son comportement dans un texte ». C'est le cas de la lexie VACHE **I.1** [*Dans le pré, des vaches broutent de l'herbe.*].

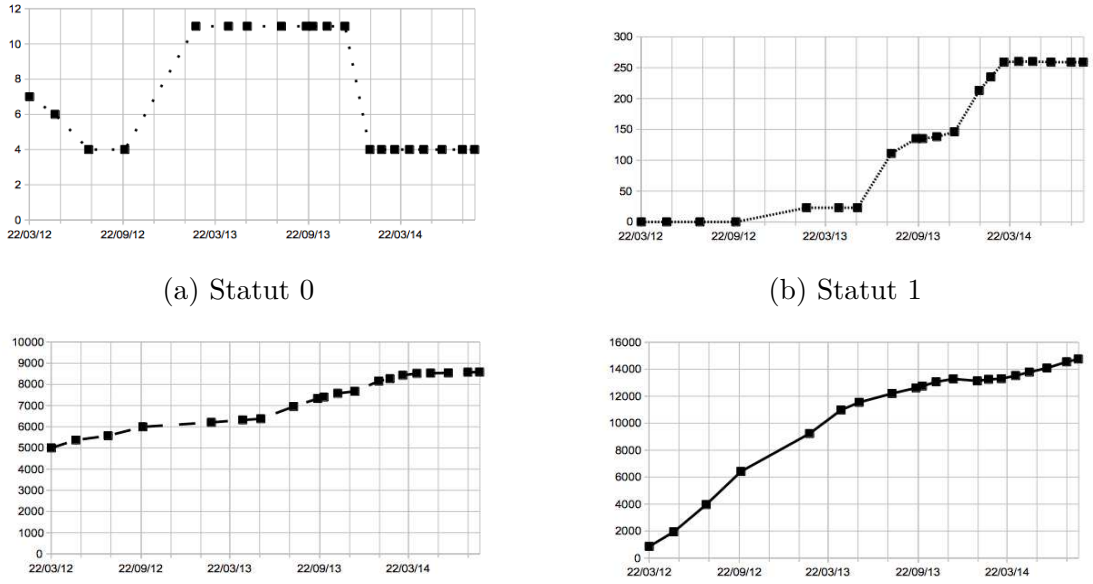
Les unités polylexématiques sont, elles, appelées *phrasèmes*. Leurs descriptions lexicographiques contiennent l'ensemble des lexies qu'elles incluent formellement. Le RL-fr en comporte deux types : des *locutions*, comme la lexie ⌜ PLANCHER DES VACHES ⌝, qui sont des unités lexicales et des expressions phraséologiques non lexicalisées, beaucoup plus marginales, comme le cliché linguistique *Comment ça va ?* Selon Polguère (2008), « une **locution** est une entité de la langue apparentée au lexème : chaque locution de la langue est structurée autour d'un sens exprimable par un ensemble de syntagmes figés, sémantiquement non compositionnels, que seule distingue la flexion ». Le modèle de leur description lexicographique est donc apparenté à celui des lexèmes. Les expressions phraséologiques non lexicalisées sont des phrasèmes compositionnels et leurs descriptions lexicographiques sont moins complexes. Dans la suite de notre travail, nous désignerons fréquemment l'ensemble des sommets comme étant des lexies. Il s'agit là d'une approximation visant à simplifier notre propos.

L'ensemble des lexies est enregistré dans une table de la base de données lexicales du RL-fr. Cette table comporte un identifiant unique pour chacune d'entre elles, l'identifiant unique du vocable auquel elle appartient, un numéro d'acception et des informations utiles au travail lexicographique. Parmi celles-ci figurent la date de dernière mise à jour et l'identité du lexicographe qui l'a effectuée. Un statut est également attribué à chaque entrée, qui prend une valeur entière comprise entre 3 et 0. Une lexie de statut 3 est une lexie qui n'a pas encore été travaillée. Le statut 2 indique une lexie en cours de traitement, le 1 une lexie en cours de vérification et le 0 une lexie dont la description est complétée. Ces trois derniers statuts ne sont pas attribués dans un ordre chronologique immuable. L'ajout d'une fonctionnalité à l'éditeur amènera, par exemple, toutes les lexies de statuts 0 à retrouver un statut 2. Le 14 août 2014, le RL-fr comportait 23 599 sommets, dont 14 757 de statut 3, 8 579 de statut 2, 259 de statut 1 et 4 de statut 0.

La figure 2.1, page suivante, montre l'évolution du nombre de lexies de chaque statut depuis la création du RL-fr jusqu'à cette date[2]. Nous y observons que le nombre de lexies de statut 0 évolue très peu pour l'instant. Il a d'ailleurs fortement diminué depuis le début de l'année 2014, où il est passé à 4 lexies. Le nombre de lexies de statut 1, pour sa part, évolue par paliers. Il a notamment connu un pic en juillet 2013, au moment de l'intégration d'une nouvelle fonctionnalité, permettant d'associer une forme propostionnelle[3] à chaque lexie. Il s'est stabilisé depuis le mois d'avril 2014. Bien que la lexie soit l'unité centrale du travail effectué, les lexicographes ne les soumettent à vérification que lorsqu'ils estiment avoir terminé l'ensemble des lexies d'un vocable. Or, le mois d'avril correspond au début de la

[2]Bien que le développement du RL-fr ait débuté en juin 2011, la fonctionnalité permettant de tisser des liens de fonctions lexicales n'a été disponible qu'au début de l'année 2012. C'est pourquoi les statistiques que nous fournissons débutent à cette période.

[3]Les formes propositionnelles, en tant qu'élément de description, sont présentées dans la section 2.2.4. Elles rendent compte des structures prédicatives des lexies.




(a) Statut 0

(b) Statut 1

(c) Statut 2

(d) Statut 3

FIG. 2.1 : Évolution du nombre de lexies par statut

systématisation d'une tâche lexicographique consistant à déterminer l'unité lexicale de base de chaque vocable d'ores et déjà créé et à en affiner la description lexicographique. La description des autres lexies de ces vocables n'étant pas suffisamment avancée, ces lexies ne sont pas immédiatement soumises à vérification. Le nombre de lexies de statut 2 et 3 évolue de façon régulière. La progression rapide du nombre de lexies de statut 3 lors de la première année correspond à la phase de constitution de la nomenclature directement induite.

### 2.2.2 Arcs relationnels typés

Nous nous sommes contentée jusqu'à présent d'énoncer que les relations encodées dans le RL-fr sont orientées et typées et que la majorité d'entre elles sont des liens sémantico-syntaxiques de fonctions lexicales Sens-Texte. Trois autres types de relations sont également présentes, elles aussi orientées et typées. Il s'agit de liens de copolysémie entre les lexies d'un même vocable, de liens d'inclusion formelle entre les phrasèmes et les lexies qui les composent et de liens d'inclusion sémantique définitionnelle entre une lexie et celles utilisées dans sa paraphrase définitonnelle. La présente section présente ces quatre types de liens et leur répartition dans la ressource.

#### 2.2.2.1 Liens de fonctions lexicales

Les fonctions lexicales, que nous appellerons désormais FL, permettent de rendre compte de la combinatoire lexicale des lexies (Mel'čuk et al., 1995, p.125-152). Elles sont disponibles dans l'éditeur lexicographique associé au RL-fr depuis le début de l'année 2012.

Selon (Mel'čuk et al., 1995, p.126), « une **fonction lexicale** [=FL] est une fonction au sens mathématique ». Elle se note traditionnellement :

$$\mathbf{F}(\textit{lexie 1}) = \{\textit{lexie 2}, \textit{lexie 3}, ...\}$$

L'ensemble de lexies {*lexie 2*, *lexie 3*, ...} est alors appelé *valeur d'application* de la FL **F** à son *argument*, ou *mot-clé*, la *lexie 1*.

Les FL s'expriment traditionnellement sous forme de formules, telles que **Magn** encodant la relation entre une lexie et ses cooccurrents d'intensification. Afin d'être compréhensible en dehors du cadre théorique Sens-Texte, chaque FL est associée à une ou plusieurs *gloses de vulgarisation* (Popovic, 2013 ; Jactel, 2013). Il s'agit de paraphrases linguistiques permettant d'énoncer en langage courant la relation établie entre l'argument d'une FL et sa valeur d'application. Ainsi, la FL **Magn** est associée à la glose « volumineux » dans le cas particulier où elle a pour argument $\text{NEZ}_{\text{Adj}}\,\mathbf{I}$ — $\mathbf{Magn}(\textit{nez}_{Adj}\,\mathbf{I}) = \{\textit{gros}_{Adj}\,\mathbf{I}\}$.

Les liens de FL, pour leur part, mettent en relation les lexies deux à deux. Le nombre de liens ayant pour source une lexie dépend donc à la fois du nombre de FL utilisées pour la décrire et du nombre d'éléments que comportent leurs valeurs d'application. Ainsi, si, comme en (1), la valeur d'application de la FL **Magn** contient une seule lexie, il existe un seul lien. En revanche si, comme en (2), elle comprend plus d'une lexie, il existe autant de liens que de lexies contenues dans cet ensemble.

**(1)** $\mathbf{Magn}(\textit{coma}) = \{\textit{profond}_{Adj}\,\mathbf{II}\}$

**(2)** $\mathbf{Magn}(\textit{aboyer}\,\mathbf{I}) = \{\textit{furieusement}\,\mathbf{I}\,;\, \textit{férocement}\}$

À partir de maintenant, nous désignerons l'ensemble des liens correspondant à des applications de FL dont une lexie est l'argument comme étant ses liens *sortants*. À l'inverse, nous parlerons de liens *entrants* pour désigner l'ensemble des liens correspondant à des applications de FL dont elle est un élément de la valeur.

Nous n'aborderons pas, ici, certains aspects des FL. Par exemple, les règles d'ordonnancement des lexies dans les valeurs d'application ne seront pas détaillées. Nous nous concentrons plutôt sur les aspects pris en considération lors de nos traitements automatiques. Nous renvoyons le lecteur à Jousse (2010) pour une présentation fine des FL en tant que système organisé.

### Des relations paradigmatiques et syntagmatiques

Il est possible de considérer l'organisation des FL, et par conséquent des liens de FL, selon plusieurs critères. En premier lieu, elles se répartissent en deux grandes classes : celles servant à encoder des relations paradigmatiques et celles servant à encoder des relations syntagmatiques.

Selon Mel'čuk et al. (1995), les FL paradigmatiques servent à spécifier l'ensemble de toutes les possibilités dans le même « paradigme » sémantique. Le 14 août 2014, sur les 779 FL disponibles dans la base de données lexicales du RL-fr, seules 297

étaient enregistrées comme étant paradigmatiques. Elles correspondaient cependant à 35 157 liens sur les 42 779 liens de FL en présence, soit plus de 82% et les dix FL les plus fréquentes, présentées dans le tableau 2.2, étaient toutes paradigmatiques[4].

| Liens | FL | Relation lexicale | Exemple |
|---|---|---|---|
| 9134 | **Syn**$_{\cap}$ | synonymie à intersection | RADIN → AVARE |
| 4236 | **Syn** | synonymie exacte | W.-C. **a** → TOILETTES **III.1a** |
| 1711 | **S$_0$** | dérivation syntaxique nominale | ABOYER **I** → ABOIEMENT **I** |
| 1558 | **Syn**$_{\subset}$ | synonymie moins riche | LESSIVE **II.1** → NETTOYANT$_N$ |
| 1348 | **S$_1$** | dérivation sémantique nominale de 1$^{er}$ actant | FLÂNER → FLÂNEUSE |
| 1110 | **Syn**$_{\supset}$ | synonymie plus riche | NETTOYANT$_N$ → LESSIVE **II.1** |
| 1071 | **V$_0$** | dérivation syntaxique verbale | ABOIEMENT **I** → ABOYER **I** |
| 1043 | **Anti** | antonymie exacte | NATUREL → ARTIFICIEL |
| 961 | **Syn**$_{\subset^{sex}}$ | synonymie moins riche relative au sexe | IMPÉRATRICE **I** → EMPEREUR |
| 960 | **Syn**$_{\supset^{sex}}$ | synonymie plus riche relative au sexe | EMPEREUR → IMPÉRATRICE **I** |

TAB. 2.2 : Dix fonctions lexicales les plus fréquentes

Les FL syntagmatiques, pour leur part, servent à spécifier « les cooccurrents [*d'une lexie*] dont la combinatoire n'est déterminée ni par leur sémantisme ni par leurs propriétés syntaxiques ». Le 14 août 2014, elles représentaient près de 62% des FL disponibles dans la base de données et 7 622 liens. La plus fréquemment utilisée était la FL d'intensification **Magn**, avec 606 liens. La seconde était la FL **Real$_1$**, avec 419 liens. Cette dernière sert à relier certaines lexies nominales aux verbes de réalisation qui ont pour sujets grammaticaux leurs premiers actants, par exemple PIANO **I** et JOUER **IV.1**. Au total, les FL syntagmatiques typaient environ 14% des arcs du RL-fr.

### Des relations standard et non standard

En second lieu, il est possible de considérer la répartition des FL selon leur statut lexicographique. En effet, la Lexicographie Explicative et Combinatoire distingue les FL selon leur « degré de standardisation » et leur complexité (Polguère, 2007). De plus, étant donné que le RL-fr est en cours d'élaboration, certaines FL ont été proposées et sont en attente de validation.

[4]La première colonne du tableau fournit le nombre d'arcs du RL-fr typés à l'aide de chacune de ces FL. Les deux dernières FL de ce tableau, **Syn**$_{\subset}^{\textbf{sex}}$ et **Syn**$_{\supset}^{\textbf{sex}}$, sont un ajout récent. Elles sont associées au FL **Fem** et **Masc**, voir à ce sujet Delaite et Polguère (2013).

Selon Jousse (2010), une FL est considérée comme *standard* si elle répond à trois critères, qu'elle nomme *cardinalité*, *diversité* et *universalité*. Le critère de cardinalité impose qu'une FL accepte un nombre important d'unités lexicales comme argument et qu'elle soit la source de nombreux liens. Ainsi, le 14 août 2014, la FL d'intensification **Magn** avait pour argument 210 lexies différentes du RL-fr, donnant lieu à 606 liens. Le critère de diversité, pour sa part, impose que les arguments d'une FL ne relèvent pas tous d'un même champ sémantique et que ses valeurs d'application varient en terme de lexies. Parmi les 210 lexies différentes servant d'argument à la FL **Magn** le 14 août 2014, 185 étaient associées à une étiquette sémantique[5]. Il s'agissait de 88 étiquettes différentes, telles que crime de sang, véhicule ou encore sensation. De plus, les 606 liens dont elle était le type avaient pour cibles 350 lexies différentes. Le critère d'universalité, pour sa part, impose que la FL encode une relation lexicale présente dans toutes les langues.

Toujours selon Jousse (2010), « les FL semi-standard sont, en quelque sorte, des extensions des FL standard permettant d'ajouter une composante de sens qui n'est pas prise en compte dans les FL standard dont elles sont dérivées ». D'après nos observations, les FL semi-standard ne respectent pas le critère de cardinalité. Ainsi, le 14 août 2014, la FL **réputation Bon** ne servait de type qu'à 19 liens ayant pour source six lexies et pour cible 11 lexies distinctes. Seules cinq d'entre elles étaient associées à des étiquettes sémantiques.

Les FL *localement standard* contreviennent au critère d'universalité. Elles sont donc, dans le cas du RL-fr, locales au français. En théorie, les FL standard et semi-standard peuvent être locales. Cependant, le 14 août 2014, le RL-fr ne comportait que des FL localement standard. Ces FL étaient au nombre de quatre et étaient très peu utilisées. Elles n'avaient pour source que cinq lexies et n'en avaient pour cibles que six.

Les FL *non standard* ne se conforment à aucun des critères énoncés. Elles permettent cependant de rendre compte de dérivations sémantiques ou de collocations. Ainsi, le 14 août 2014, la FL **Excrément de** ∼ servait à typer 15 liens du RL-fr, à partir de cinq lexies source, vers sept lexies différentes. Ces cibles étaient en grande majorité associées à l'étiquette sémantique substance corporelle.

Un dernier « degré de standardisation » est utilisé dans le RL-fr. Il regroupe les FL de *lexicalisation de régime*. Ces FL indiquent des réalisations lexicalisées d'éléments du régime syntaxique de l'argument auquel elles s'appliquent. Par exemple, la lexie SAVON **I.b** a deux actants syntaxiques, le premier étant l'utilisateur, celui qui nettoie, le second le patient, ce qui est nettoyé. La FL **$2='corps'** permet de relier cette lexie à son collocatif TOILETTE **I.1a**, pour former la collocation spécifique au cas de figure où ce qui est nettoyé à l'aide d'un savon est un corps : *savon de toilette*. Le 14 août 2014, la base de données du RL-fr comportait 17 FL de lexicalisation de régime. Seules 14 d'entre elles étaient utilisées pour typer des liens. Ces derniers reliaient 18 sources à 23 cibles.

[5]Les étiquettes sémantiques sont présentées dans la section 2.2.4.

Le critère de complexité des FL s'applique aux FL standard et semi-standard, qu'elles soient locales ou non. Il permet de distinguer les FL simples, telles que la FL d'intensification **Magn**, des FL résultant de la combinaison de plusieurs FL simples, telle que la FL de nominalisation par une unité minimale régulière $\mathbf{SingS_0}$ ou encore la FL **Bon + Magn** qui combine le sens de la FL d'approbation subjective du locuteur et celui de celle d'intensification.

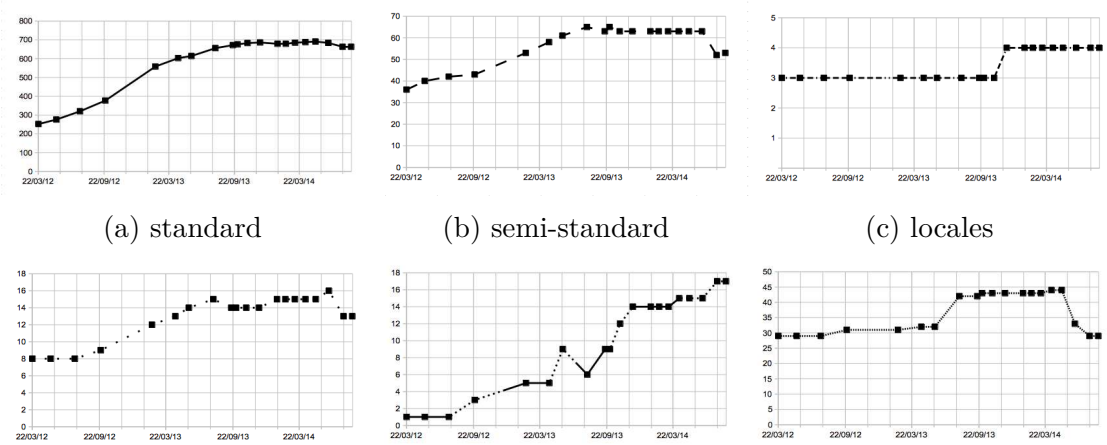
Le 14 août 2014, la base de données du RL-fr comptait 779 FL. Le tableau 2.3 présente leur répartition en fonction des différents statuts que nous venons d'énoncer. Nous pouvons y observer que la grande majorité de FL encodées sont standard.

| Statut | Nbr. | Exemple FL | Exemple lien |
|---|---|---|---|
| standard simples | 273 | **Magn** | VENT **I.1** → VIOLENT$_{\text{Adj}}$ **II** |
| standard complexes | 390 | $\mathbf{CausFunc_0}$ | DOUCHE **II.1** → INSTALLER **II.1** |
| semi-standard simples | 36 | **réputation Bon** | AVOCATE **I** → RENOMMÉ |
| semi-standard complexes | 17 | **déplacement** $\mathbf{S_0Fact_0}$ | AVION **II** → VOL$^{1}$ **2b** |
| localement standard complexes | 4 | $\mathbf{De\_nouveauFunc_0}$ | CHEVEU **II.a** → REPOUSSER$^{2}$ |
| standard simples proposées | 7 | **Fem** | COQ → POULE$^{1}$ **I** |
| standard complexes proposées | 17 | $\mathbf{Enun_1Real_1}$ | HOLD-UP **I** → HAUT LES MAINS ! |
| semi-standard simples proposées | 5 | **fonctionnel Mero** | MOTO **I** → MOTEUR$_{\text{N}}$ |
| non standard | 13 | **Excrément de** $\sim$ | CHEVAL **I.1a** → CROTTIN |
| lexicalisation de régime | 17 | **$2='corps'** | SAVON **I.b** → *de toilettes* **I.1a** |

TAB. 2.3 : Répartition des FL par statut.

Les lignes doubles du tableau 2.3 permettent de distinguer six grandes classes de statuts. La figure 2.2 montre l'évolution du nombre de FL pour chacune de ces classes depuis la création du RL-fr. Nous y observons que le nombre de FL n'est pas en constante augmentation, mais qu'il connait des vagues de vérification et de « nettoyage ». Le nombre total de FL est ainsi passé de 832 en mai 2014 à 778 en juillet de la même année. Cette diminution a affecté l'ensemble des classes de FL, à l'exception de la classe très réduite des FL locales et de la classe des lexicalisations de régime.

Les lexicographes prévoit de poursuivre la révision de l'ensemble des FL tout au long du développement du RL-fr. L'un des objsectif visé par cette révision est la diminution de la granularité du système mis en œuvre dans l'encodage des relations syntactixo-sémantique, en évitant notamment un foisonnement des FL non standard.



(a) standard (b) semi-standard (c) locales

(d) non standard (e) lexicalisation (f) proposées

FIG. 2.2 : Évolution du nombre de FL par statut

### Des familles de FL

En dernier lieu, il est possible de considérer la répartition des FL en fonction de leur famille. Ces familles correspondent à des types de relations et sont utilisées pour optimiser l'interface de travail des lexicographes. Elles leur permettent d'accéder rapidement à une FL souhaitée. Par exemple, si un lexicographe souhaite encoder une relation de synonymie, il sélectionnera la famille **Syn** et accédera à l'ensemble des dix FL suivantes :

**Synonymie exacte :** $\mathsf{Syn}(pull) = \{pull\text{-}over\}$ ;

**à intersection de sens :** $\mathsf{Syn}_{\cap}(pull) = \{sweat,\ sweat\text{-}shirt\}$ ;

**plus riche :** $\mathsf{Syn}_{\supset}(fixer) = \{clouer\,\mathbf{I}\}$ ;

**moins riche :** $\mathsf{Syn}_{\subset}(clouer\,\mathbf{I}) = \{fixer\}$ ;

**plus riche relative au sexe :** $\mathsf{Syn}_{\supset}^{\mathsf{sex}}(sénateur) = \{sénatrice\}$ ;

**moins riche relative au sexe :** $\mathsf{Syn}_{\subset}^{\mathsf{sex}}(sénatrice) = \{sénateur\}$ ;

**Hyponymie :** $\mathsf{Hypo}(félins_N) = \{chat^1\,\mathbf{I.a}\}$ ;

**Confer :** $\mathsf{Cf}(abeille\,\mathbf{I.a}) = \{bourdon\,;\ frelon\,;\ guêpe\}$ ;

**Nom du chef d'un ensemble régulier :** $\mathsf{CapMult}(abeille\,\mathbf{I.a}) = \{reine\,\mathbf{II}\}$ ;

**Variation formelle :** $\mathsf{variante\ formelle}\ \mathsf{Syn}(kilogramme) = \{kilo\}$.

Le 14 août 2014, la base de données du RL-fr comptait 108 familles. La majorité d'entre elles, comme la famille **Syn**, comportait soit des FL paradigmatiques, soit des FL syntagmatiques. Il existait cependant 16 familles « mixtes ». La famille **Son**, par exemple, était alors composée des 3 FL suivantes, dont seule la première est syntagmatique :

**$\sim$ émettre un son : Son**$(bouilloire) = \{siffler$ ***II***$,\ chanter$ ***II.1***$\}$ ;

**son émis par $\sim$ : S$_0$Son**$(chatte$ ***I***$) = \{miaulement\}$ ;

**son émis par $\sim$ (onomatopée) : EnunSon**$(abeille$ ***I.a***$) = \{$« *Bzz* »$\}$.

Les familles de FL regroupent des FL dont le « degré de standardisation » et la complexité varient fortement. Les FL complexes sont à rechercher dans la famille correspondant à leur dernier élément. C'est généralement cet élément qui sera énoncé en premier lors de l'expression vulgarisée d'une telle FL ou de la relation dont un arc ainsi typé rend compte entre deux lexies. Les formules de FL s'interprètent ainsi de droite à gauche.

#### 2.2.2.2 Liens de copolysémie

Les liens de copolysémie sont disponibles dans l'éditeur lexicographique associé au RL-fr et dans sa base de données depuis décembre 2013. Ils permettent de rendre compte de la structure polysémique des vocables en encodant les liens existants entre les lexies qui y sont regroupées. La majorité de ces liens encodent des relations sémantiques. Comme nous l'avons déjà souligné, ces liens sont orientés. Chacun a pour cible une lexie dont le sens est dérivé de celle qui en est la source[6]. Les lexies d'un vocable ne forment donc pas une clique, dans laquelle l'ensemble des lexies seraient reliées deux à deux par des liens de copolysémie. Comme le montre la figure 2.3, qui présente les liens tissés entre les lexies du vocable CHAT[1], la structure d'un vocable est une structure d'arbre.

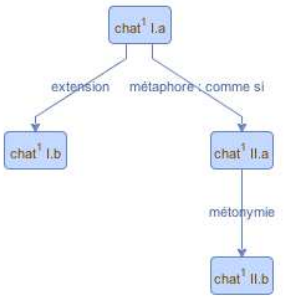



FIG. 2.3 : Polysémie du vocable CHAT[1]

Dans ce cas précis, l'unité lexicale de base du vocable est la lexie CHAT[1] **I.a** [*Christelle a adopté deux chats*]. La seconde acception, CHAT[1] **I.b** [*Le puma est le plus gros chat du*

[6]Le lien de codérivation fait exception à cette règle. Le sens des deux lexies qu'il relie est dérivé du sens de lexies d'un autre vocable. Par exemple, les lexies QUÉBECOIS$_N$ **1** [*Les Québécois ront élu un Premier Ministre.*] et QUÉBECOIS$_N$ **2** [*À Québec, on parle le québécois.*] dérivent du sens de lexies du vocable QUÉBEC.

*monde.*] a un sens plus large. Elle est dérivée de l'unité de base par extension. La troisième, CHAT[1] **II.a** [*« C'est toi le chat... »*], est métaphorique. Elle dénote un participant spécifique dans un jeu de cour de récréation. La dernière est dérivée de la troisième par métonymie et dénote le jeu lui-même, CHAT[1] **II.b** [*Ce matin, à la récré, on a joué à chat*].

#### 2.2.2.3 Types et sous-types

Le 14 août 2014, le RL-fr comptait 4 505 liens de copolysémie, correspondant à deux types de relations non sémantiques et à sept types de relations sémantiques. Nous les listons ici, accompagnés d'exemples de liens, dans l'ordre où ils apparaissent dans l'éditeur lexicographique.

**causation :** POLLUER **1** [*Les nitrates polluent.*]

→ POLLUER **2** [*Les voitures polluent.*] ;

**résultat :** SALIR **1** [*Il a sali sa veste.*]

→ SALIR **2** [*La boue salit ses bottes.*] ;

**conversion :** COMMENCER **I.2a** [*Le repas commence par une soupe.*]

→ COMMENCER **I.2b** [*Une soupe commence le repas.*] ;

**spécialisation :** MOYENNE **I.1** [*La moyenne des prix est de 12€.*]

→ MOYENNE **I.2** [*Il a eu la moyenne dans toutes les matières*] ;

**extension :** CHAT[1] **I.a** [*Christelle a adopté deux chats*]

→ CHAT[1] **I.b** [*Le puma est le plus gros chat du monde.*] ;

**intersection**[7] **:** FLOTTE **I.1** [*La flotte anglaise mouille au large de Brest.*]

→ FLOTTE **II** [*La flotte n'a pas cessé de tomber depuis deux heures.*] ;

**métonymie :** CHAT[1] **II.a** [*« C'est toi le chat... »*]

→ CHAT[1] **II.b** [*Ce matin, à la récré, on a joué à chat*] ;

**métaphore$_{\text{comme si}}$ :** CHAT[1] **I.a** [*Christelle a adopté deux chats*]

→ CHAT[1] **II.a** [*« C'est toi le chat... »*] ;

**codérivation :** QUÉBECOIS$_{\text{N}}$ **1** [*Les Québécois ront élu un Premier Ministre.*]

→ QUÉBECOIS$_{\text{N}}$ **1** [*À Québec, on parle le québécois.*] ;

L'un des types que nous venons de citer est accompagné d'un sous-type, **comme si**. Il s'agit du sous-type le plus général disponible pour cette relation. Le 14 août 2014, il était associé à 58% des liens métaphoriques. Huit autres sous-types sont disponibles pour préciser la nature de certains liens de copolysémie. Le sous-type **sous-sens** est disponible pour la spécialisation, les sous-types **ensemble de** et **partie de** pour la métonymie et **forme**, **utilisation**, **fonction**, **fonctionnement** et **comportement** viennent en complément de **comme si** pour la métaphore.

[7] Le type de relation encodée ici correspond à de la quasi-homonymie.

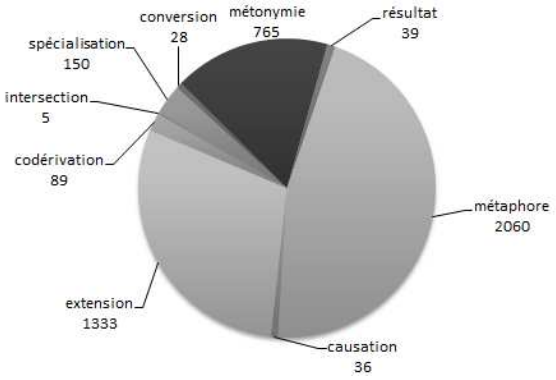



FIG. 2.4 : Répartition des liens de copolysémie par types

La figure 2.4 montre la répartition des liens de copolysémie par type, en date du 14 août 2014. Nous y observons que la grande majorité de liens de copolysémie encodés est de type métaphore, métonymie et extension.

#### 2.2.2.4 Liens d'inclusion formelle

Les liens d'inclusion formelle sont disponibles dans l'éditeur lexicographique associé au RL-fr et dans sa base de données depuis septembre 2012. Il s'agit de liens asémantiques, reliant les phrasèmes — essentiellement des locutions — aux lexies qui les composent. Ils sont orientés des phrasèmes vers leurs composants et ne se subdivisent en aucun sous-type. La locution nominale ⌜ PLANCHER DES VACHES ⌝ est ainsi reliée à deux lexies, $\text{PLANCHER}_{\text{N}}$ et VACHE **I.a**.

Le 14 août 2014, nous pouvions comptabiliser 5 798 liens d'inclusion formelle, ce qui représentait près de 11% des arcs. La majorité d'entre eux avaient été encodés dans le cadre du travail de master de Marie-Sophie Pausé (Pausé, 2014), qui poursuit actuellement sa recherche par un doctorat portant sur la structure lexico-sémantique des locutions.

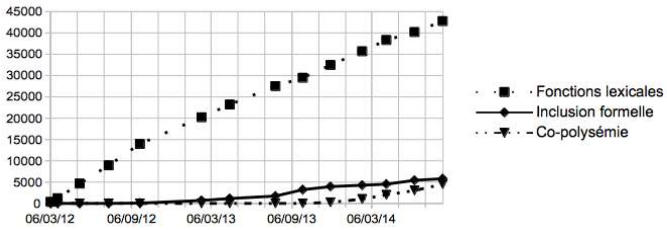



FIG. 2.5 : Évolution du nombre d'arcs

La figure 2.5 montre l'évolution du nombre de liens d'inclusion formelle depuis la création du RL-fr, comparativement au nombre de liens de FL et de copolysémie.

#### 2.2.2.5 Liens d'inclusion sémantique définitionnelle

Selon Polguère (2014b), les systèmes lexicaux comportent un quatrième type d'arcs relationnels. Il s'agit de liens d'inclusion sémantique définitionnelle, qui relient chaque lexie à l'ensemble des lexies utilisées dans sa paraphrase définitionnelle. La figure 2.6 fournit un exemple d'une telle paraphrase, pour la lexie HOLD-UP I [*Un hold-up a eu lieu hier dans une bijouterie du quartier. Deux individus sont en garde à vue*]. Nous pouvons y voir que quatre liens ont été encodés. Chacun de ces liens a pour source la lexie HOLD-UP I et pour cible l'une des lexies ATTAQUE I.2, ⌜ LIEU PUBLIC ⌝, ARME et VOL[2] I.

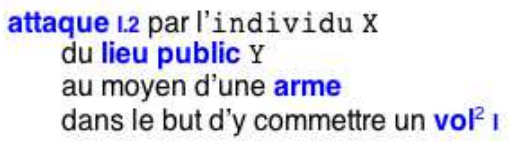



FIG. 2.6 : Paraphrase définitionnelle de HOLD-UP I

Tout comme les paraphrases définitionnelles, ce type de liens n'est actuellement présent dans le RL-fr que de façon expérimentale. Ainsi, le 14 août 2014, il n'en comportait que seize. Le développement des fonctionnalités de l'éditeur lexicographique permettant leur implantation est en cours.

### 2.2.3 Arcs relationnels et analogie

Nous venons de présenter les différents types d'arcs encodés dans le RL-fr. Comme nous l'avons vu, ces arcs rendent compte de relations entre sommets lexicaux et ces relations sont en grande partie sémantiques.

Si nous considérons, à la suite de Lepage (2003), que l'analogie est une « conformité de rapport entre objets du même type » et que les arcs du RL-fr établissent des rapports entre sommets lexicaux, il semble naturel de considérer deux couples de sommets partageant un même type d'arcs comme étant analogues.

La présente section propose un premier examen de cette hypothèse, basée sur la construction de proportions analogiques. Elle pose les bases de la mise en place d'une mesure de similarité de Relations entre couples de lexies dans le cadre d'une exploration du RL-fr par raisonnement analogique.

#### 2.2.3.1 Liens de FL et analogie

Le cas des liens de FL est particulièrement intéressant dans le cadre de l'hypothèse examinée ici. En effet, Mel'čuk et al. (1995) définissent l'application des FL standard comme des proportions. Ils illustrent leur propos à l'aide de la FL **Magn** et de son application à PLEURER et à PLUIE. Selon eux, « l'expression *comme une Madeleine* remplit par rapport à PLEURER (à peu près) le même rôle que l'adjectif préposé *grosse* par rapport à PLUIE » et ils ajoutent que « pour être une FL, une dépendance lexicale doit [...] donner lieu à un grand nombre de proportions de ce

genre ».

Comme nous l'avons vu dans le premier chapitre, section 1.2, les proportions peuvent être considérées comme l'expression mathématique de relations d'analogie. L'illustration ci-dessus doit donc pouvoir être reformulée de la manière suivante : « *comme une Madeleine* **est à** PLEURER **ce que** *grosse* **est à** PLUIE ». Le rapport que cette analogie amène à énoncer correspond alors à la FL elle-même : « l'expression lexicale de son intensification ».

Selon cette logique, l'ensemble des couples de lexies reliées par des liens de même FL donnent lieu à des analogies. Ainsi, les couples de lexies en lien de nominalisation $\mathbf{S_0}$ (DORMIR**I.1a** [*J'ai dormi deux heures hier après-midi.*], SOMME$_{\text{N,masc}}$ [*Elle pique un somme sur sa chaise.*]) et (ILLUSTRER**II** [*Il est utile de choisir un bon exemple pour illustrer ses propos.*], ILLUSTRATION**I** [*L'histoire de la pomme de Newton est une illustration du concept de gravité.*]) permettent de formuler que « SOMME$_{\text{N,masc}}$ **est à** DORMIR**I.1a ce que** ILLUSTRATION**I est à** ILLUSTRER**II** ».

Nous avons vu également dans le chapitre 1, section 1.2.3, que Lepage (2003) propose de découper l'ensemble des analogies possibles pour quatre objets donnés (A, B, C et D) en trois classes de huit analogies équivalentes. Si l'analogie ci-dessus est correcte, nous pourrions donc supposer que les sept suivantes le sont aussi :

1. ILLUSTRATION**I est à** ILLUSTRER**II ce que** SOMME$_{\text{N,masc}}$ **est à** DORMIR**I.1a**
2. ILLUSTRER**II est à** ILLUSTRATION**I ce que** DORMIR**I.1a est à** SOMME$_{\text{N,masc}}$
3. DORMIR**I.1a est à** SOMME$_{\text{N,masc}}$ **ce que** ILLUSTRER**II est à** ILLUSTRATION**I**
4. SOMME$_{\text{N,masc}}$ **est à** ILLUSTRATION**I ce que** DORMIR**I.1a est à** ILLUSTRER**II**
5. ILLUSTRATION**I est à** SOMME$_{\text{N,masc}}$ **ce que** ILLUSTRER**II est à** DORMIR**I.1a**
6. DORMIR**I.1a est à** ILLUSTRER**II ce que** SOMME$_{\text{N,masc}}$ **est à** ILLUSTRATION**I**
7. ILLUSTRER**II est à** DORMIR**I.1a ce que** ILLUSTRATION**I est à** SOMME$_{\text{N,masc}}$

Il nous est alors possible d'énoncer, sans difficulté, le rapport mis en avant par la première de ces analogies. Il correspond lui aussi à la FL $\mathbf{S_0}$. Les deux suivantes sont toutes aussi naturelles. Elles correspondent à la FL de verbalisation $\mathbf{V_0}$, connue pour être l'inverse systématique de la FL $\mathbf{S_0}$. Les quatre analogies suivantes sont en revanche problématiques. De quel rapport s'agit-il ici ? Comment peut-on énoncer « ce que SOMME$_{\text{N,masc}}$ est à ILLUSTRATION**I** » ? Et selon quelle dimension ?

Polguère (2008, p.160) définit la valeur d'application d'une FL comme étant composée d'éléments liés « *(à peu près)* de la même façon » à la lexie qui en est l'argument. Cette approximation de conformité, déjà énoncée par Mel'čuk et al. (1995), est-elle la raison de cette difficulté ? Faut-il considérer que le rapport difficile à énoncer est en réalité absent et que le cube des expressions analogiques équivalentes proposé par Lepage (2003) ne s'applique pas aux proportions entre lexies établies selon un rapport de FL ? Qu'il ne s'applique pas aux analogies entre objets particuliers que sont les lexies ? Ou encore que les FL ne permettent pas d'établir des

conformités de rapports entre lexies ?

Considérons à présent l'application de la même FL de nominalisation $\mathbf{S_0}$ aux lexies MIAULER [*Son chat miaule pour sortir.*] et ABOYER**I** [*Le chien des voisins aboie à la fenêtre*], appartenant au même champ sémantique. Il est alors possible d'énoncer l'ensemble des huit analogies suivantes :

1. MIAULEMENT **est à** MIAULER **ce que** ABOIEMENT**I** **est à** ABOYER**I**
2. ABOIEMENT**I** **est à** ABOYER**I** **ce que** MIAULEMENT **est à** MIAULER
3. MIAULER **est à** MIAULEMENT **ce que** ABOYER**I** **est à** ABOIEMENT**I**
4. ABOYER**I** **est à** ABOIEMENT**I** **ce que** MIAULER **est à** MIAULEMENT
5. MIAULEMENT **est à** ABOIEMENT**I** **ce que** MIAULER **est à** ABOYER**I**
6. ABOIEMENT**I** **est à** MIAULEMENT **ce que** ABOYER**I** **est à** MIAULER
7. MIAULER **est à** ABOYER**I** **ce que** MIAULEMENT **est à** ABOIEMENT**I**
8. ABOYER**I** **est à** MIAULER **ce que** ABOIEMENT**I** **est à** MIAULEMENT

Dans ce cas de figure, nous ne rencontrons plus les mêmes difficultés à énoncer les rapports mis en évidence. Pour les quatre premières analogies, les rapports correspondant aux FL de nominalisation et de verbalisation sont toujours disponibles. Pour les quatre dernières, il est possible d'énoncer l'un des deux rapports suivants : « son équivalent pour les lexies CHAT[1]**I.a** et CHATTE**I** » et « son équivalent pour les lexies CHIEN**I.a** et CHIENNE**I** ». Comme nous le voyons, ces deux rapports sont énoncés à partir d'autres lexies. De façon plus générale, ils correspondent à une information relative à la position des sommets du graphe lexical. Bien que les lexicographes consultés ne voient aucun intérêt à l'énonciation de tels rapports, cette observation attise notre curiosité. En ajoutant la contrainte de lexies appartenant à un même champ lexical, nous avons spécifié la nature des objets mis en relation analogique. Est-ce cela qui a permis d'établir un rapport en terme de position ou les rapports précédemment difficiles à énoncer peuvent-ils également trouver une expression dans une dimension topologique ?

Nous ne trancherons ici en faveur d'aucune des hypothèses que nous venons d'énoncer. Nous utiliserons en revanche les liens de FL entre lexies pour mesurer la similarité de Relations entre couples. Ainsi, nous considérerons, de ce point de vue, les couples (DORMIR**I.1a**, $\text{SOMME}_{\text{N,masc}}$) et (ILLUSTRER**II**, ILLUSTRATION**I**) tout aussi similaires que les couples (MIAULER, MIAULEMENT) et (ABOYER**I**, ABOIEMENT**I**). Nous garderons tout de même à l'esprit le questionnement que nous venons de soulever lors de notre exploration du RL-fr par raisonnement analogique[8].

[8]Nous reviendrons notamment sur cette question dans le chapitre 5, section 5.6.4.

#### 2.2.3.2 Copolysémie et analogie

En adaptant à la copolysémie la logique que nous venons d'appliquer aux liens de FL, nous pouvons considérer que l'ensemble des couples de lexies reliées par des liens de copolysémie de même type donnent lieu à des analogies. Ainsi, les couples de lexies (BOUCHER$_N$ **I.2** [*Un boucher expérimenté coupe la viande finement.*], BOUCHER$_N$ **I.1** [*Je vais chez le boucher acheter de la viande.*]) et (CHARCUTIER$_N$ **I.2** [*Un charcutier qualifié est un vrai professionnel de la découpe.*], CHARCUTIER$_N$ **I.1** [*Chez le charcutier, on achète jambon, saucisson et saucisses.*]), dont la première lexie dérive de la seconde par **extension** permettent de formuler que « BOUCHER$_N$ **I.2 est à** BOUCHER$_N$ **I.1 ce que** CHARCUTIER$_N$ **I.2 est à** CHARCUTIER$_N$ **I.1** ».

La granularité du typage des liens de copolysémie est cependant moins fine que celle des FL et les lexies utilisées pour construire la proportion ci-dessus partagent un nombre important d'éléments de description lexicographique[9] en plus d'une relation de dérivation sémantique similaire. Si nous comparons désormais le couple de lexies (BOUCHER$_N$ **I.2**, BOUCHER$_N$ **I.1**) au couple (CHAT$^1$ **I.b** [*Le puma est le plus gros chat du monde.*],CHAT$^1$ **I.a** [*Christelle a adopté deux chats*]), dont nous avons vu dans la section 2.2.2.2 que la première lexie dérive également de la seconde par **extension**, nous aboutissons à la formulation de l'analogie « BOUCHER$_N$ **I.2 est à** BOUCHER$_N$ **I.1 ce que** CHAT$^1$ **I.b est à** CHAT$^1$ **I.a** ». Même si l'analogie ainsi exprimée demeure correcte et qu'elle met en évidence le rapport correspondant au lien de copolysémie « une acception dérivée par extension de sens », elle semble moins convaincante. Tandis que l'analogie précédente permettait de rendre compte d'une dérivation régulière en français entre une lexie dénotant un individu dont le métier consiste à vendre quelque chose et une lexie dénotant un individu dont le métier consiste à préparer cette même chose, aucune abstraction pertinente n'est accessible à partir de celle-ci.

Ces observations rejoignent les considérations de Gentner (1983), Medin et al. (1990) et Turney (2006) sur la distinction entre analogies proches et analogies lointaines, que nous avons présenté au chapitre 1, section 1.5.4.2. Elles nous confortent dans l'idée que l'exploration automatique d'une ressource lexicale par raisonnement analogique nécessite la mise au point d'une mesure de similarité d'Attributs tout autant que d'une mesure de similarité de Relations.

De la même manière que nous l'avons annoncé pour les liens de FL, nous utiliserons les liens de copolysémie pour mesurer la similarité de Relations entre couples de lexies. Nous considérerons, de ce point de vue, que les couples (VALISE **I** [*Sa valise est prête pour partir demain.*], VALISE **II** [*J'ai des poches...non...Des valises sous les yeux.*]) et (FLÛTE **I** [*Pierre joue de la flûte et du banjo.*], FLÛTE **IV** [*Tu pourras passer acheter une flûte, à la boulangerie ?*]) sont similaires. Chacun de ces couples est composé de lexies dont la seconde a un sens dérivé de la première par métaphore de forme. De plus, comme le suggèrent les travaux de Barque (2008), Barque et Chaumartin (2009), Goossens (2011) et Sikora (2014), nous serons attentive à l'analogie entre structures polysémiques des vocables.

[9]Les éléments de description lexicographique sont présentés dans la section 2.2.4.

#### 2.2.3.3 Inclusion formelle et analogie

Contrairement aux liens de FL et de copolysémie, les liens d'inclusion formelle ne sont pas sémantiques. De plus, aucun type particulier ne leur est associé. La construction de proportions analogiques entre couples de lexies basée exclusivement sur la présence de liens d'inclusion formelle aboutirait donc à des énoncés aussi variés que « PERCHÉ **est à** ⌜CHAT PERCHÉ⌝ **ce que** MUSICAL **est à** ⌜CHAISE MUSICALE**I**⌝ », « FAUTE**I.1a est à** *Pour faute* **ce que** SANS**II.1 est à** ⌜SANS FAUTE$_{\text{Adv}}$⌝ » ou « MOUCHE**I est à** ⌜CHASSER LES MOUCHES⌝ **ce que** PLAGE**I.1a est à** ⌜VOLLEY DE PLAGE**1**⌝ ».

La prise en compte d'informations supplémentaires, telles que les structures syntaxiques des phrasèmes, la position des lexies à l'intérieur de celles-ci ou les caractéristiques grammaticales de ces dernières, permettraient sans doute d'améliorer la qualité des analogies ainsi énoncées. De telles perspectives sortent cependant du cadre de nos recherches.

Les liens d'inclusion formelle demeurant des liens relationnels, nous choisissons de ne pas les exclure a priori de toute mesure de similarité de Relations. Nous conservons ainsi la possibilité d'observer l'influence d'une mesure de similarité d'Attributs sur les couples de lexies partageant une telle Relation. La pertinence de leur prise en compte sera évaluée de manière empirique dans le cadre de nos expériences.

### 2.2.4 Descriptions lexicographiques encapsulées

Comme nous l'avons énoncé précédemment, une particularité des systèmes lexicaux est la non-atomicité de leurs sommets. Chacune des unités lexicales qu'ils comportent est associée à une description lexicographique complexe. Lors de l'exploration du RL-fr par raisonnement analogique, les éléments constitutifs de cette description pourront être considérés comme des Attributs de lexies et être utilisés pour établir des mesures de similarité.

Comme nous venons de le voir dans la section 2.2.3, nous pensons que de telles mesures de similarité permettront de spécifier la nature des objets lexicaux mis en relation analogique lors de cette exploration. Les analogies qu'il sera alors possible d'énoncer seront plus pertinentes que des analogies basées uniquement sur une mesure de similarité de Relations.

La présente section détaille l'ensemble des éléments de description lexicographiques disponibles. À des fins d'illustration, elle s'appuie sur la description des trois lexies regroupées dans le vocable ABEILLE — ABEILLE**I.a** [*Cette ruche est pleine d'abeilles.*], ABEILLE**I.b** [*Cet apiculteur est vraiment passionné par les abeilles.*] et ABEILLE**II** [*Ce fauteuil est orné d'abeilles et de fleurs de lys.*] — et sur leur visualisation dans l'éditeur lexicographique Dicet (Gader et al., 2012). Elle fournit également des informations sur l'évolution dans le temps de chacun des éléments de description depuis le début de l'année 2012.

L'éditeur Dicet propose une vue groupée de l'ensemble des lexies d'un même vocable. La figure 2.7 représente cette vue pour le vocable ABEILLE. Nous pouvons y

voir, dans la partie gauche, des informations relatives au vocable lui-même. Dans la partie en haut à droite, une visualisation de sa structure polysémique est proposée. En dessous, nous distinguons neuf onglets. Ces onglets correspondent chacun à une partie de la description des lexies.

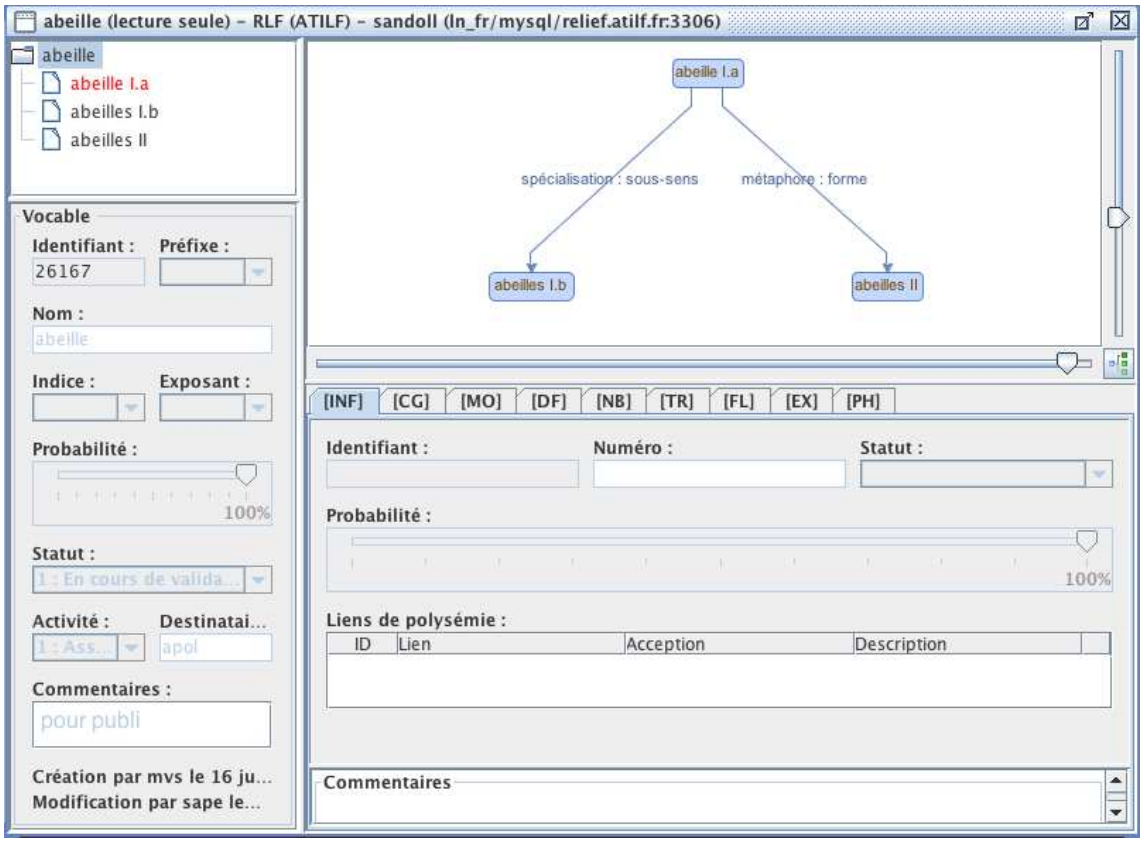



FIG. 2.7 : Vue du vocable ABEILLE dans l'éditeur Dicet

Nous parlerons ici exclusivement des onglets [CG], [MO], [DF] et [EX]. En effet, les onglets [INF] et [FL] concernent les informations relatives aux lexies et aux FL, présentées dans les sections 2.2.1 et 2.2.2.1. Les onglets [NB] et [TR], réservés à des remarques lexicographiques et au tableau de régime syntaxique ne disposent, pour l'instant, d'aucune fonctionnalité autre qu'un champ de commentaires et ne sont pas exploitables automatiquement. Enfin, l'onglet [PH] permet aux lexicographes de visualiser l'ensemble des phrasèmes formés à partir de la lexie sélectionnée. Il correspond donc aux relations présentées en 2.2.2.4, parcourues en sens inverse.

#### 2.2.4.1 Appartenance à un vocable

Comme nous le voyons sur la figure 2.7, chaque lexie est associée à un vocable. L'ensemble des vocables est enregistré dans une table de la base de données du RL-fr. Cette table comporte les informations que nous pouvons voir ici : un identifiant unique, un éventuel préfixe, une forme graphique ou nom, d'éventuels indice et exposant, un indice de confiance ou probabilité, un statut, une « activité » associée à un destinataire, d'éventuels commentaires, ainsi que les dates de création et de der-

nière modification, associées à l'identité des lexicographes qui ont réalisé ces actions.

Le préfixe est utilisé dans le cas de verbes pronominaux, il prend la forme *se* ou *s'*. L'indice est utilisé pour distinguer différents vocables ayant la même forme graphique, mais se distinguant en terme de partie du discours et/ou de genre. Par exemple, le RL-fr compte actuellement trois vocables ayant pour forme *mélomane*, l'adjectif MÉLOMANE$_{\text{Adj}}$, le nom féminin MÉLOMANE$_{\text{N, fém}}$ et le nom masculin MÉLOMANE$_{\text{N, masc}}$. L'exposant est utilisé pour distinguer les vocables homonymes et prend la forme d'un chiffre. Le statut des vocables est comparable à celui des lexies que nous avons présenté dans la section 2.2.1. Il prend une valeur entière comprise entre 3 et 1. Le statut 3 indique un vocable encore non travaillé. Ce statut n'implique pas que toutes les lexies que le vocable regroupe soient elles-mêmes non travaillées. Cela signifie en revanche qu'il est possible que certaines acceptions du vocable soient absentes ou que son unité lexicale de base ne soit pas encore dégagée. Le statut 2 indique un vocable en cours de traitement, le 1 un vocable en cours de validation. Les vocables de statut 1 ne comportent que des lexies de statut 1, elles-mêmes en cours de validation. L'activité et le destinataire permettent de savoir si un lexicographe travaille actuellement sur le vocable considéré. Le vocable ABEILLE est ainsi assigné à **apol**. Si un autre lexicographe souhaite y apporter une modification, il se doit de le contacter pour lui en faire part.

Le 14 août 2014, le RL-fr comptait 15 656 vocables, dont 121 de statut 1 et 5 463 de statut 2. Un peu plus de 74% de l'ensemble des vocables étaient alors monosémiques. Cette valeur tombait cependant à 55% pour les seuls vocables en cours de traitement et à 19% pour ceux en cours de validation. La figure 2.8 montre l'évolution du nombre global de vocables depuis le début de l'année 2012. Elle est accompagnée de l'évolution du nombre de lexies de statut 3, en pointillés gris. Ces deux évolutions sont proches l'une de l'autre. Elles connaissent notamment le même pic de croissance au cours de la première année, lors de la phase de constitution de la nomenclature directement induite.

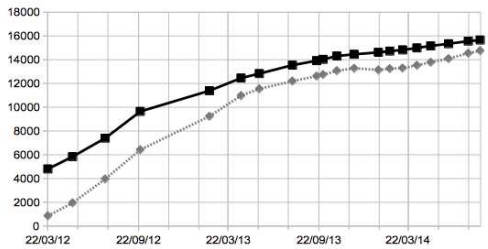



FIG. 2.8 : Évolution du nombre de vocables

Le fait, pour deux lexies, d'appartenir à un même vocable est une similarité d'Attributs de peu d'importance dans un modèle qui contient des liens de copolysémie. Le sens porté par ces liens est en effet plus précis. Nous exploiterons cependant cet élément de description au cours de nos premières expériences, réalisées avant leur implémentation.

#### 2.2.4.2 Caractéristiques grammaticales

L'onglet [CG] permet aux lexicographes d'associer une lexie et l'ensemble des caractéristiques grammaticales disponibles dans la base de données du RL-fr qu'ils jugent pertinentes dans son cas. La figure 2.9 montre cet onglet dans le cas de la lexie ABEILLES **I.b**.

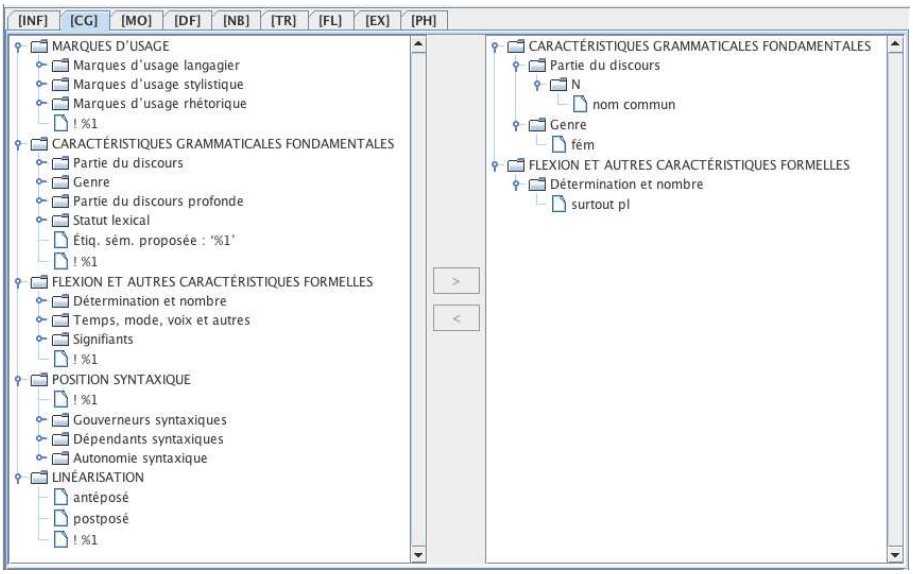



FIG. 2.9 : Caractéristiques grammaticales de ABEILLES **I.b**

La partie gauche offre un aperçu de l'ensemble des caractéristiques disponibles, sous forme d'un menu hiérarchique. Les lexicographes sélectionnent dans ce menu les caractéristiques qu'ils attribuent à la lexie en cours de traitement. Elles constituent le premier élément d'information encodé pour chacune d'entre elles. A minima, chaque lexie doit être associée à une partie du discours et, s'il s'agit d'une lexie nominale, à un genre.

La partie de droite montre les caractéristiques auxquelles la lexie ABEILLES **I.b** a été associée : la partie du discours `nom commun`, le genre féminin `fém` et la caractéristique flexionnelle `surtout pl`, qui spécifie qu'elle s'emploie surtout au pluriel.

Comme nous l'avons déjà évoqué, le nombre de caractéristiques grammaticales disponibles dans la base de données évolue en fonction des besoins des lexicographes. Alors qu'au début de l'année 2012, il était de 103 ; en août 2014, 172 caractéristiques étaient utilisées pour réaliser 44 221 associations. Il est d'ailleurs intéressant de noter que ces associations posent peu de problèmes aux lexicographes et que 44 026 d'entre elles disposaient d'un indice de confiance de 100%. Seules 33 lexies n'étaient impliquées dans aucune de ces associations, dont 26 étaient des lexies destinées à

être supprimées[10]. La figure 2.10 montre l'évolution du nombre de caractéristiques dans le temps. Nous pouvons y voir qu'après une période de stabilité entre septembre 2012 et août 2013, il a connu une certaine croissance et qu'il s'est à nouveau stabilisé à partir du mois de mai 2014.

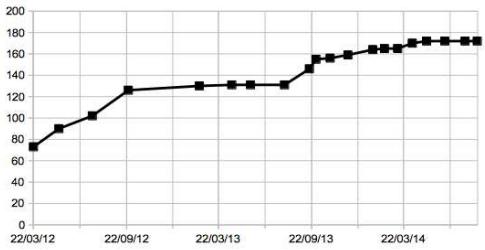



FIG. 2.10 : Évolution du nombre de caractéristiques grammaticales

Le fait, pour deux lexies, de partager les mêmes caractéristiques grammaticales semble a priori intéressant à exploiter dans une mesure de similarité d'Attributs. Nous verons cependant que leur diversité et leur granularité peut amener à considérer comme étant différentes des lexies au fonctionnement analogue. Nous réfléchirons donc à une manière optimale de les exploiter.

#### 2.2.4.3 Morphologie

L'onglet [MO] permet aux lexicographes d'associer à chaque lexie une table flexionnelle prototypique pour générer l'ensemble des flexions qui lui correspondent[11]. C'est aussi l'occasion pour les lexicographes d'en spécifier la forme de nommage. Ainsi, la figure 2.11 montre que la lexie ABEILLES **I.b** — dont nous venons de voir qu'elle est surtout employée au pluriel — a pour forme de nommage *abeilles*. Elle est associée à la table flexionnelle prototypique `chat`, avec un indice de confiance de 100%.

Chaque fois qu'une table flexionnelle prototypique est associée à l'unité lexicale de base d'un vocable, la même table est associée automatiquement aux autres lexies regroupées dans ce vocable. Un indice de confiance de 50% est alors attribué à cette association. Les lexicographes utilisent l'indice de 60% pour distinguer les cas dont ils souhaitent discuter rapidement et l'indice de 90% pour les incertitudes qui devront être levées au moment de la validation des lexies, ou les cas pour lesquels la table nécessaire n'a pas encore été créée. Ils laissent alors un commentaire dans le champ prévu à cet effet. Ainsi, le 14 août 2014, 52 associations avaient un indice de 50%, 22 un indice de 60% et 87 un indice de 90%.

[10]Il arrive que les lexicographes aient à supprimer une acception, soit parce qu'ils révisent la polysémie d'un vocable, soit parce qu'elle a été créée accidentellement, par « faute de clic ». Afin d'éviter des suppressions malencontreuses, cette opération ne peut être réalisée que par les lexicographes qui ont un statut particuliers d'administrateur, comme **apol**. Afin d'être facilement repérées, les lexies à supprimer sont numérotées **0**.

[11]Nous vous invitons à la lecture de Gader et al. (2014a) pour une présentation détaillée de l'implémentation de cette fonctionnalité dans l'éditeur Dicet.

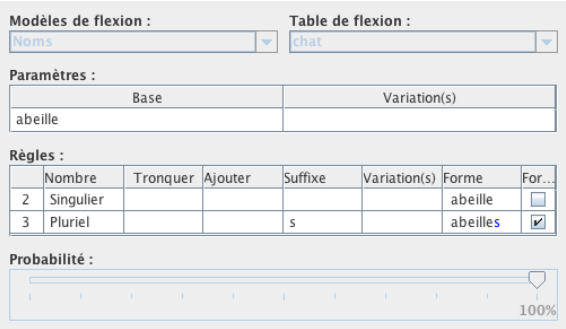



FIG. 2.11 : Morphologie de ABEILLES **I.b**

Certaines lexies, pour lesquelles il existe une variante graphique, font l'objet de plusieurs associations. C'est le cas de la lexie ABÎMÉ, visible dans la figure 2.12. C'est alors la première association qui sert de forme de nommage de référence.

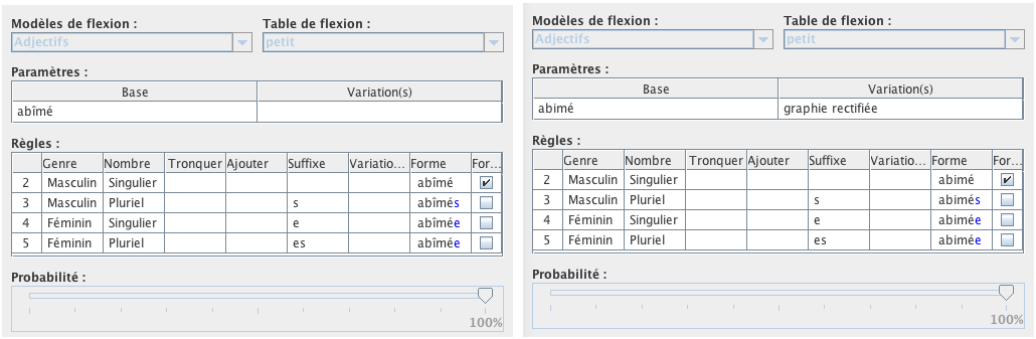
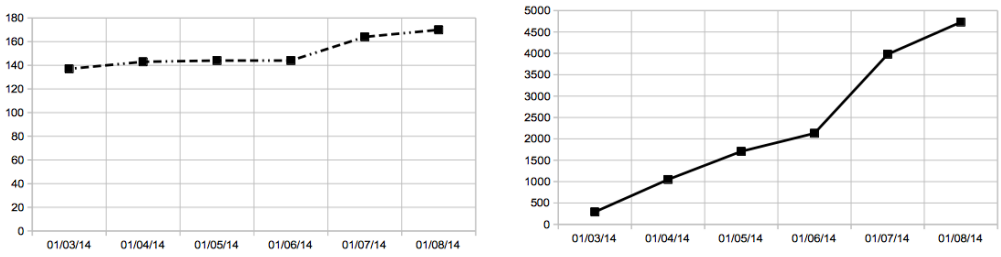



FIG. 2.12 : Morphologie de ABÎMÉ

L'encodage des informations flexionnelles est disponible dans l'éditeur lexicographique depuis le mois de mars 2014. Depuis cette date, il est effectué par deux lexicographes spécialisées. Elles réalisent cette tâche pour l'ensemble des lexies nominales, verbales et adjectivales de la nomenclature, en la parcourant par ordre alphabétique. Le 14 août 2014, 170 tables prototypiques étaient disponibles et associées à 4 602 lexies, par l'intermédiaire de 4 729 associations. À peu près 20% de l'ensemble des lexies disposait donc de cet élément d'information. Tout comme les caractéristiques grammaticales, les tables prototypiques disponibles dans la base de données évoluent en fonction des besoins rencontrés. La figure 2.13 montre l'évolution de leur nombre, ainsi que de celui des associations table-lexie pour cette période de six mois.

Dans le cadre d'une exploration du RL-fr par raisonnement analogique, le fait que deux lexies partagent ou non la même table flexionnelle prototypique ne nous paraît pas être une similarité d'Attributs pertinente. Cet élément de description ne sera donc pas exploité lors de nos expériences[12].

[12]Ces informations pourraient cependant s'avérer utiles dans le cadre d'une annotation de corpus à partir du RL-fr.




(a) Tables prototypiques (b) Associations table-lexie

FIG. 2.13 : Évolution du nombre d'informations morphologiques

#### 2.2.4.4 Définitions

L'onglet [DF] permet aux lexicographes d'associer aux lexies les éléments de description relatifs à leur définition, autrement dit leur description sémantique. Ces éléments se divisent en trois parties : étiquette sémantique, forme propositionnelle et paraphrase définitionnelle. Comme nous l'avons précédemment évoqué, les fonctionnalités liées à cette dernière ne sont pas encore toutes opérationnelles. Certaines paraphrases sont cependant d'ores et déjà présentes. Elles ont en partie été entrées dans le champ commentaires, comme nous pouvons le voir dans la figure 2.14, page 64, pour la lexie ABEILLE **I.a**.

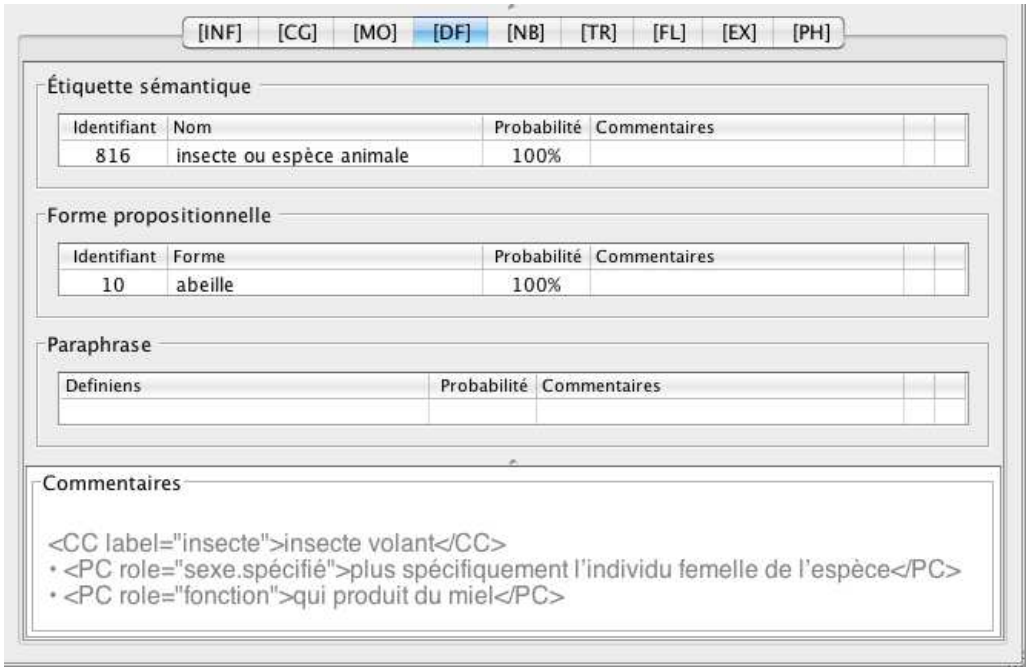



FIG. 2.14 : Définition de ABEILLE **I.a**

Conformément à la Lexicologie Explicative et Combinatoire, les définitions du RL-fr sont des *définitions analytiques*. Comme le rappelle Polguère (2011a), une telle définition, aussi appelée définition par genre prochain et différences spécifiques, « est une modélisation du sens d'une lexie L qui possède les trois propriétés fondamentales

suivantes : (i) c'est une paraphrase exacte de L, (ii) elle est constituée de lexies sémantiquement plus simples que L, (iii) elle est structurée en une composante centrale, le *genre prochain*, qui est la paraphrase minimale de L, et un ensemble de composantes périphériques, les *différences spécifiques*, qui particularisent le sens de L relativement à son genre prochain et aux autres lexies ayant le même genre prochain ».

### Étiquettes sémantiques

Comme nous pouvons le voir sur la figure 2.14, la première partie de la description sémantique d'une lexie correspond à son étiquette. Il s'agit de la forme normalisée du genre prochain de sa définition analytique. Elle est sélectionnée par les lexicographes dans une ontologie de classes, présentée en détail par Polguère (2011a). Cette ontologie est purement linguistique et ne fait référence à aucun concept extérieur à la langue.

Les instances des classes sont les étiquettes sémantiques à proprement parler. Ce sont elles qui sont associées aux lexies. Contrairement aux classes, qui sont toutes nominales, les étiquettes peuvent également être verbales, adjectivales, adverbiales et clausatives. En tant que forme normalisée du genre prochain des définitions, elles doivent en effet être en correspondance avec les parties du discours des lexies.

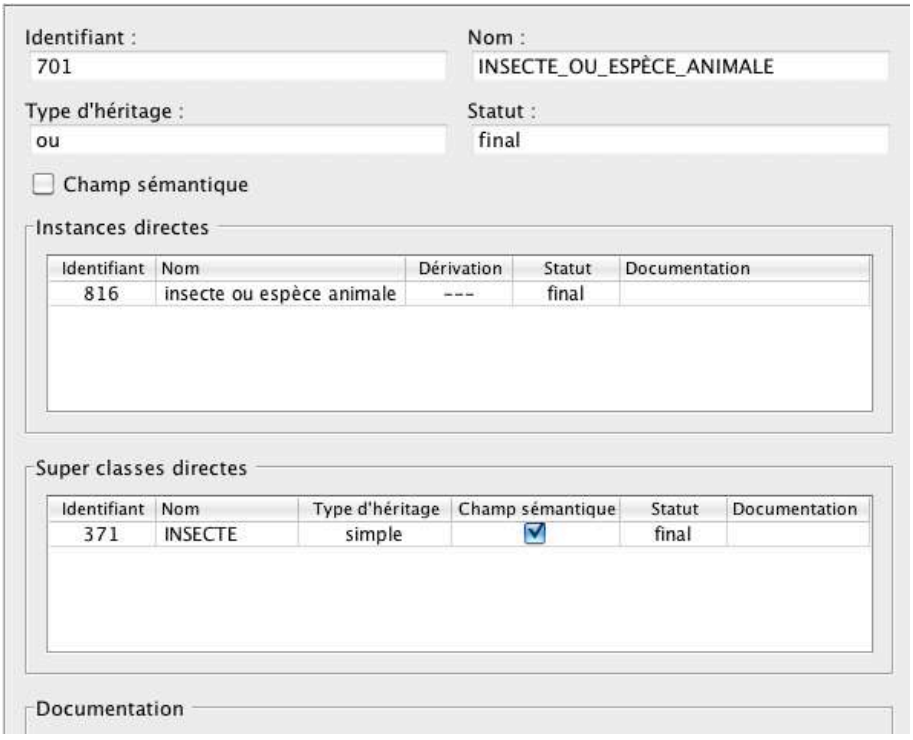



FIG. 2.15 : Classe INSECTE_OU_ESPÈCE_ANIMALE

La figure 2.15 présente les détails de la classe INSECTE_OU_ESPÈCE_ANIMALE à laquelle est rattachée la lexie ABEILLE **I.a**, par l'intermédiaire de l'instance insecte ou espèce animale. Comme nous pouvons le voir, chaque classe dispose d'un identifiant unique, d'un nom, d'un type d'héritage, d'un statut, d'un ensemble d'instances,

d'une ou plusieurs classes mères et d'une éventuelle documentation. Certaines classes se voient également associer à un attribut **champ sémantique**.

Les types d'héritage sont au nombre de trois. L'héritage *simple* est le cas général. Le 14 août 2014, il était associé à 531 des 593 classes repertoriées dans la base de données asssociée au RL-fr. Les classes qui disposent de ce type sont rattachées à une seule classe mère. Les cas associés aux deux autres types sont des cas d'héritage multiple et sont rattachés à deux classes mères. L'héritage *ou* correspond aux étiquettes disjonctives exclusives. C'est le cas de l'étiquette insecte ou espèce animale. Les lexies qui disposent de cette étiquette dénotent soit une instance particulière d'insecte, soit une espèce animale, mais jamais les deux à la fois. Le 14 août 2014, 24 classes repertoriées disposaient de ce type d'héritage. L'héritage *et/ou* correspond aux étiquettes disjonctives inclusives. C'est le cas des étiquettes de la classe ÉNONCÉ, présentée dans la figure 2.16. Les lexies nominales qui disposent de telles étiquettes dénotent soit un contenu informationnel, soit son contenant. Elles peuvent éventuellement dénoter les deux à la fois, comme la lexie QUESTION, dans l'exemple *Regardez bien les tics présidentiels durant la question, qui apparaît dans sa traduction anglaise en bas de l'écran*[13].Le 14 août 2014, 38 classes repertoriées disposaient de ce type d'héritage.

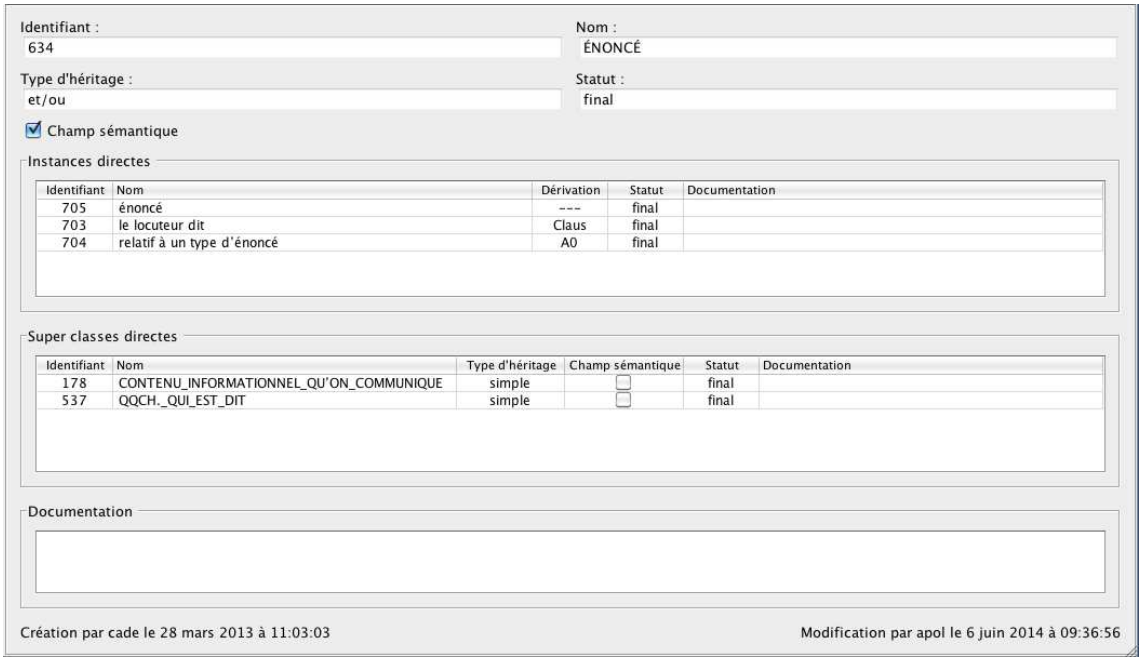



FIG. 2.16 : Classe ÉNONCÉ

Les statuts sont au nombre de deux, ils permettent de distinguer les classes simplement proposées de celles d'ores et déjà validées. Le 14 août 2014, seules 12 classes n'étaient pas validées.

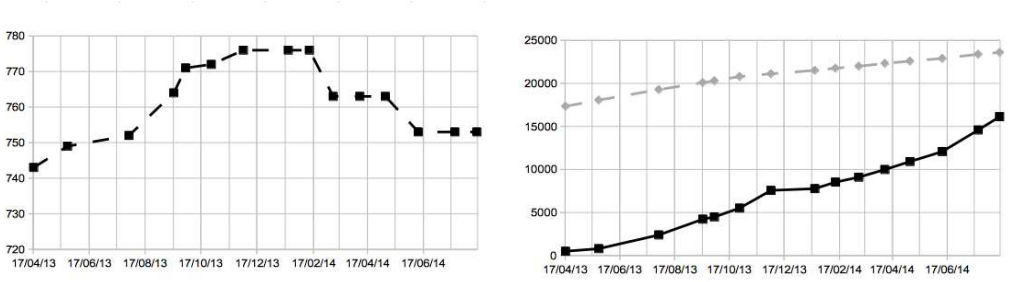
Comme nous pouvons le voir dans la figure 2.16, l'étiquette nominale d'une classe est considérée comme son instance de base, dont les autres dérivent. Ainsi, l'étiquette

[13]Cet exemple est emprunté à Milicevic et Polguère (2010), qui discutent du cas particuliers de l'ambivalence sémantique des noms de communication langagière.

relatif à un type d'énoncé est une dérivation adjectivale de l'étiquette énoncé.

L'attribut **champ sémantique** sert à distinguer les classes qui jouent un rôle important dans l'organisation sémantique du lexique. C'est le cas de la classe ÉNONCÉ, mais non de la classe INSECTE_OU_ESPÈCE_ANIMALE. Le 14 août 2014, la première des deux était la mère de trois autres classes et les instances de ces quatres classes réunies servaient à étiqueter 181 lexies. La seconde, en revanche, n'avait aucune classe fille et ses instances ne servaient à étiqueter que deux lexies. Elle jouait donc bien un rôle de moindre importance dans l'organisation du RL-fr.

Comme l'ensemble des éléments de description, l'ontologie de classes et d'étiquettes sémantiques se transforme au fur et à mesure des besoins rencontrés par les lexicographes. Le 14 août 2014, 753 étiquettes sémantiques étaient disponibles et servaient à étiqueter 16 128 lexies, soit près de 70% d'entre elles. Cet étiquetage est toutefois particulièrement délicat et seules 40% de ces associations disposaient d'un indice de confiance de 100%. La figure 2.17 montre l'évolution du nombre d'étiquettes et d'association étiquette-lexie depuis l'implémentation des fonctionnalités permettant leur encodage jusqu'à cette date. Le nombre d'associations y est comparé au nombre de lexies, en gris. Nous pouvons y voir que le nombre d'étiquettes n'est pas en constante augmentation, mais qu'il a diminué au cours du premier semestre 2014. Tout comme nous l'avons évoqué pour la morphologie, deux lexicographes sont spécialisées dans la gestion de l'ontologie et effectuent des vérifications, des ajouts et des suppressions de manière régulière. Elles mènent actuellement un travail visant à réduire la granularité de la classification.




(a) Étiquettes sémantiques (b) Associations étiquette-lexie

FIG. 2.17 : Évolution du nombre d'étiquettes sémantiques

Le fait, pour deux lexies, de partager une même étiquette sémantique nous semble, de prime abord, être une similarité d'Attributs pertinente pour l'exploration par raisonnement analogique. La figure 2.17 montre cependant que leur couverture du RL-fr ne dépasse les 50% que depuis le mois de juin 2014. De plus, comme nous venons de le voir, les indices de confiance des associations étiquette-lexie sont relativement faibles. Indépendamment de ces considérations numériques, l'exploitation d'une telle similarité oriente le rapport analogique établi entre lexies en le situant dans la dimension de l'organisation du lexique en champs sémantiques. Nous serons donc amenée à l'écarter lorsque cette dimension ne nous intéressera pas.

**Formes propositionnelles (Definiendum)**

La seconde partie de la définition d'une lexie correspond à sa forme propositionnelle, aussi appelée *definiendum* et désormais FP. Elle rend compte de sa structure actancielle. ABEILLE**I.a** étant une lexie non prédicative, sa FP est minimale et reprend simplement sa forme graphique `abeille`. En revanche, celle de la lexie ABEILLE**I.b** rend compte de l'existence d'une position actancielle et prend la forme `abeille élevée par X=1`.

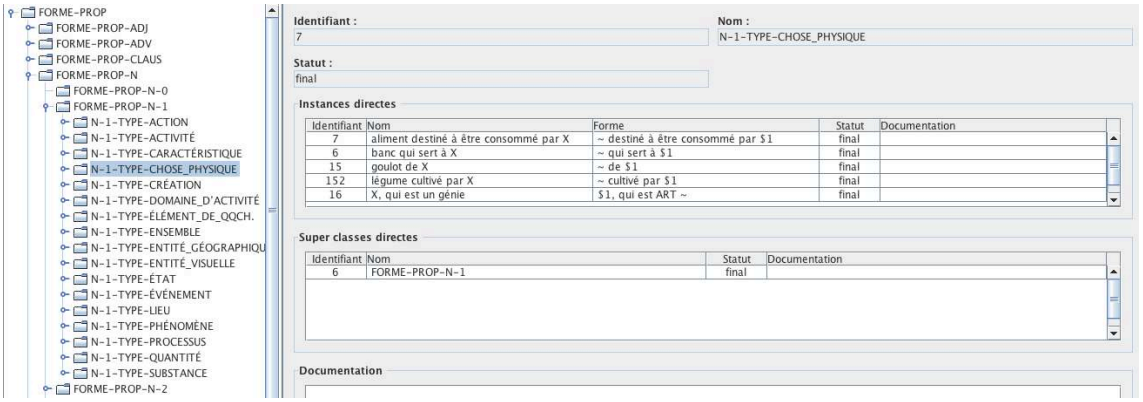
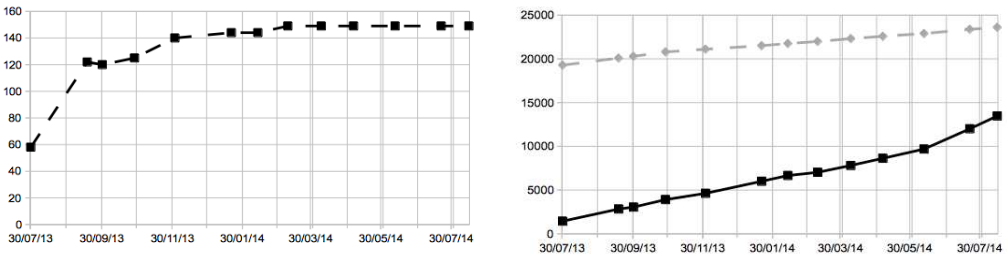



FIG. 2.18 : Forme propositionnelle `N-1-TYPE-CHOSE_PHYSIQUE`

Les lexicographes sélectionnent la FP qu'ils souhaitent associer à une lexie dans une ressource organisée en classes, tout comme l'ontologie d'étiquettes sémantiques. Les FP à proprement parler sont des instances de ces classes. La partie gauche de la figure 2.18 donne un aperçu de cette hiérarchie. La partie droite fournit les informations relatives à la classe des lexies nominales à un actant, dénotant des choses physiques. La FP de la lexie ABEILLE**I.b** appartient à cette classe. Elle correspond à l'instance enregistrée sous le nom `légume cultivé par X`.

Les fonctionnalités permettant d'encoder cet élément de description ont été implémentées dans l'éditeur Dicet en juillet 2013. Le 14 août 2014, 149 FP distinctes étaient disponibles, réparties en 73 classes et associées à 13 471 lexies. Près de 95% de ces associations disposaient d'un indice de confiance de 100%. La figure 2.19 montre l'évolution du nombre de FP et d'associations FP-lexie entre ces deux dates. Le nombre d'associations y est comparé au nombre de lexies, en gris. Nous y observons que le nombre de FP disponibles n'a pas évolué depuis mars 2014. Cela ne signifie pas pour autant qu'il a atteint son maximum. Ainsi, seules 29 FP sont actuellement disponibles pour les lexies verbales, contre 82 pour les lexies nominales.

Tout comme les étiquettes sémantiques, les FP nous semblent, de prime abord, permettre d'établir des similarités d'Attributs intéressantes entre lexies. Leur implémentation tardive et leur faible couverture, moins d'un tiers des lexies associées à une FP jusqu'au mois d'avril 2014, nous a cependant dissuadée de les exploiter lors de nos expériences.




(a) Formes propositionnelles (b) Associations FP-lexie

FIG. 2.19 : Évolution du nombre de FP

### Paraphrase définitionnelle (Definiens)

La troisième partie de la description sémantique d'une lexie correspond à sa paraphrase définitionnelle, ou *definiens*. Comme nous l'avons déjà évoqué, celle-ci se décompose en une composante centrale, le genre prochain, et en composantes périphériques. La figure 2.20, reprend la figure 2.6 de la page 54, qui fournit un exemple d'une telle définition, pour la lexie HOLD-UP I.

**attaque** I.2 par l'individu X
du **lieu public** Y
au moyen d'une **arme**
dans le but d'y commettre un **vol**[2] I

FIG. 2.20 : Paraphrase définitionnelle de HOLD-UP I

Sans entrer dans les détails de la constitution de cette paraphrase, nous pouvons voir que la première ligne correspond au genre prochain de la lexie. Les trois suivantes représentent chacune l'une de ses composantes périphériques.

Le 14 août 2014, le RL-fr ne comportait que sept paraphrases définitionnelles. Celles-ci n'ont donc pas été prises en compte dans nos expériences.

#### 2.2.4.5 Exemples lexicographiques

L'onglet [EX] permet aux lexicographes d'associer un ensemble d'exemples d'emploi à chaque lexie. La figure 2.21, présentée page suivante, montre cet onglet dans le cas de la lexie ABEILLES II.

Comme le décrit Lux-Pogodalla (2014), les exemples sélectionnés par les lexicographes sont majoritairement issus d'un regroupement de corpus dit *corpus-réservoir*. Les corpus qui s'y trouvent sont constitués de textes littéraires, d'articles journalistiques et d'extraits du Web.

Les lexicographes interrogent les textes littéraires à travers l'interface de la base

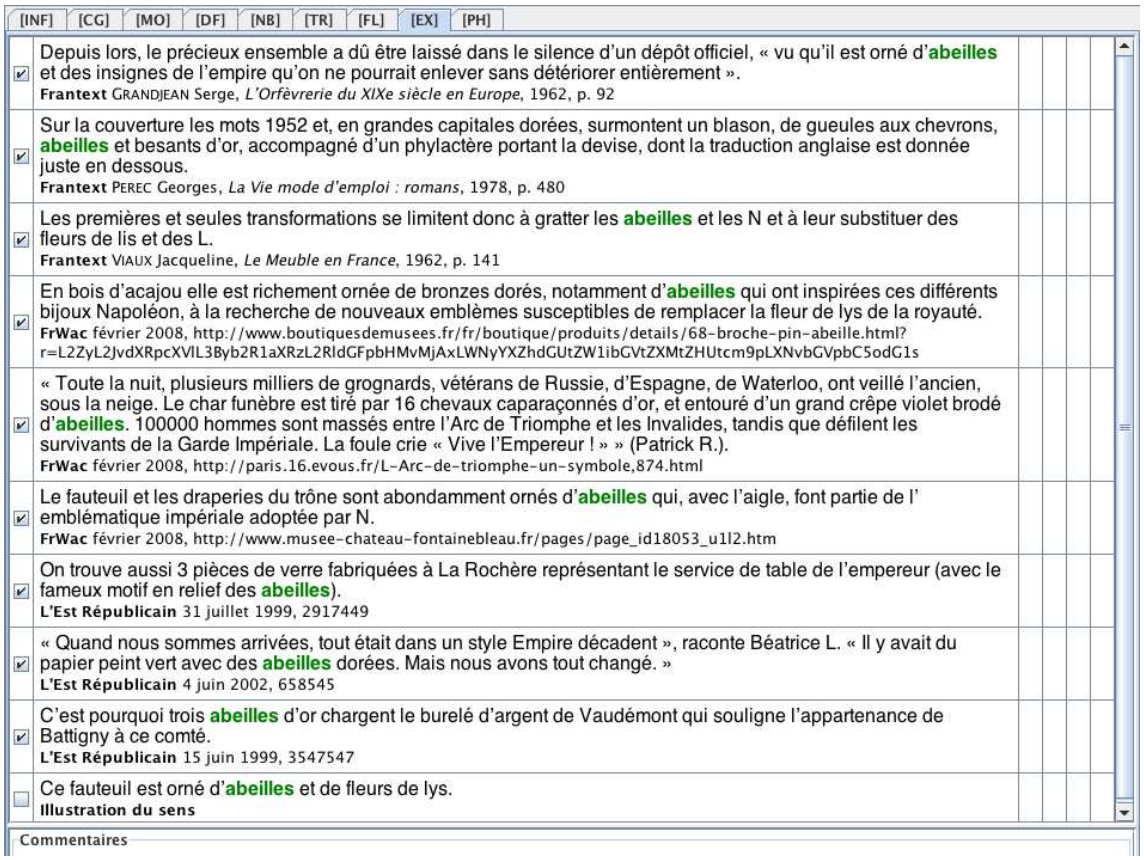



FIG. 2.21 : Exemples d'emploi de ABEILLES II

de données textuelles Frantext[14]. Ils selectionnent par avance l'ensemble des textes datant d'après 1950. Cette pré-sélection ne leur assure cependant pas que les textes aient été **rédigés** après 1950. En effet, les dates disponibles dans cette base de données ne correspondent pas aux dates d'écriture des ouvrages, mais à leurs dates d'édition. Il est donc important que les lexicographes soient vigilants, au cours de la tâche de description des lexies, tout autant que pendant celle de vérification. Les articles de presse sont issus du corpus Est Républicain, disponible sur le site du Centre National de Ressources Textuelles[15]. Il est composé d'articles du journal régional *L'Est Républicain* écrits en 1999, 2002 et 2003. La partie Web du corpus-réservoir est constituée du corpus FrWac[16] (Baroni et al., 2009). Elle comporte des textes collectés automatiquement en 2008 sur le domaine **.fr**.

Les lexicographes ont pour consigne d'associer neuf exemples d'emploi à chaque lexie qu'ils décrivent et au moins un exemple aux lexies qu'ils créent en tissant des liens. Ces exemples doivent présenter la variété du comportement syntaxique de la lexie et la réalisation de ses actants sémantiques, sans pour autant être exhaustifs. Un exemple d'illustration, le plus concis possible, doit également être associé aux lexies en cours de traitement. L'ensemble de ces exemples doit comporter un vocabulaire simple et des occurrences non ambigues de la lexie qu'il illustre. Au besoin,

[14] `http ://www.frantext.fr`.
[15] `http ://cnrtl.fr/corpus/estrepublicain/`.
[16] `http ://nl.ijs.si/noske/wacs.cgi/corp_info ?corpname=frwac`.

des exemples extraits du Web ou de publications tout venant, de supports audiovisuels, de conversations ou de courriers personnels peuvent également être utilisés. Certains peuvent également être fabriqués, mais le recours à ces derniers doit être exceptionnel.

Chaque exemple utilisé est enregistré dans une table de la base de données du RL-fr, accompagné d'un ensemble d'informations bibliographiques. Si des modifications lui ont été apportées, comme le remplacement d'un pronom par un groupe nominal ou l'anonymisation d'un nom propre, elles sont spécifiées dans un champ commentaires. De telles modifications doivent cependant être évitées. Les lexicographes préféreront choisir un exemple en dehors de leur corpus-réservoir plutôt que d'y avoir recours. La figure 2.22 montre le résultat de cet enregistrement pour l'un des exemples associés à la lexie ABEILLES **II**. Nous pouvons y voir que ce même exemple est associé à une seconde lexie.


ID : 3696
Classe : Citations de corpus journalistiques hors Frantext
Source : L'Est Républicain
Statut : En attente de validation (1)
Citation
« Quand nous sommes arrivées, tout était dans un style Empire décadent », raconte Béatrice L. « Il y avait du papier peint vert avec des abeilles dorées. Mais nous avons tout changé. »
Référence
**L'Est Républicain** 4 juin 2002, 658545
Occurrences

| | |
|---|---|
| abeilles II | « Quand nous sommes arrivées, tout était dans un style Empire décadent », raconte Béatrice L. « Il y avait du papier peint vert avec des **abeilles** dorées. Mais nous avons tout changé. » |
| vert I.1a | « Quand nous sommes arrivées, tout était dans un style Empire décadent », raconte Béatrice L. « Il y avait du papier peint **vert** avec des abeilles dorées. Mais nous avons tout changé. » |

Création par sape le 17 janvier 2013 à 16:57:43
Modification par eju le 9 juillet 2013 à 16:22:15

FIG. 2.22 : Exemple extrait de l'*Est Républicain*

L'ensemble des exemples enregistrés constitue une collection d'exemples lexicographiques, dont au moins une lexie peut-être considérée comme étant annotée par le RL-fr. Cette lexie dispose donc potentiellement d'autant de couches d'annotation distinctes que d'éléments de descriptions que nous avons vu jusqu'à présent. Elle dispose, par exemple, d'une étiquette sémantique, qui permet d'envisager l'interrogation de la collection selon la dimension des champs sémantiques. Les lexicographes sont invités à consulter cette collection avant le corpus-réservoir lors de la recherche de nouveaux exemples d'emploi. Cette démarche permet une désambiguïsation progressive de cette ressource encapsulée dans le RL-fr. Elle est favorisée par la mise à disposition d'un export quotidien, rendu compatible avec l'outil de textométrie TXM (Heiden et al., 2010). Son concordancier permet alors aux lexicographes de bénéficier de toutes les subtilités de la syntaxe CQL[17] pour interroger cette collec-

[17]Les lexicographes utilisent également cette syntaxe dans les interfaces à leur disposition pour interroger les corpus FrWac et Est Républicain.

tion.

Les fonctionnalités permettant d'enregistrer les exemples dans la base de données du RL-fr et de les associer à des lexies sont disponibles depuis le début de l'année 2013. Un certain nombre d'exemples a été auparavant enregistré dans un champ commentaire de la table de lexies. Ils ne sont pas exploitables automatiquement, mais sont réintégrés plus formellement au fur et à mesure du travail lexicographique actuel. Le 14 août 2014, 25 444 exemples servaient à illustrer 19 002 lexies, par l'intermédaire de 37 264 associations. Cependant, près de 76% de ces lexies n'étaient associées qu'à un seul exemple. L'ensemble des associations disposait d'un indice de confiance de 100%. L'export rendu disponible sous TXM correspondant, qui exclut les simples illustrations de sens, comportait 25 235 exemples, 1 068 587 occurrences de formes, 88 313 formes distinctes et 39 578 lemmes[18]. Près de 47% de ces exemples étaient extraits de la base de données textuelles Frantext, 20% du FrWac, 16% de l'Est Républicain et 13% du Web tout venant. La figure 2.23 présente l'évolution du nombre d'exemples, en trait continu, et d'association exemple-lexie, en pointillés. Nous y observons que l'écart grandit progressivement entre ces deux nombres, ce qui signifie que de plus en plus d'exemples sont utilisés pour illustrer plusieurs lexies distinctes.

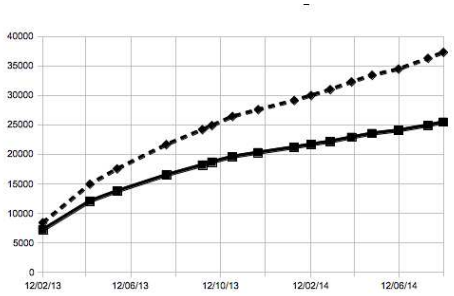



FIG. 2.23 : Évolution du nombre d'exemples

Le 14 août 2014, près de 69% des exemples ne servaient à illustrer qu'une seule lexie, tandis que quatre d'entre eux servaient à en illustrer une dizaine. Le tableau 2.4 montre la répartition des exemples en fonction du nombre de lexies qui leur étaient associées à cette date.

| Exemples | 4 | 3 | 9 | 21 | 65 | 175 | 586 | 1 667 | 5 454 | 17 460 |
|---|---|---|---|---|---|---|---|---|---|---|
| Lexies | 10 | 9 | 8 | 7 | 6 | 5 | 4 | 3 | 2 | 1 |

TAB. 2.4 : Répartitions des exemples par nombre de lexies

Le fait, pour deux lexies, de partager un même exemple est de toute évidence trop restrictif pour être exploité dans une mesure de similarité d'Attributs. Il est cependant envisageable de se servir de cet élément de description pour établir une

[18]L'étiquetage morpho-syntaxique permettant le calcul de lemmes est effectué à l'aide de TreeTagger (Schmid, 1994, 1995).

mesure de similarité de Relations entre lexies. La disparité du nombre d'exemples disponibles, de 1 à 62 pour une lexie donnée, nous a cependant amenée à laisser cette question de côté lors de nos expériences. Une éventuelle exploitation de ces exemples nécessiterait de s'interroger sur les différents degrés d'annotation possibles. Est-il pertinent de comparer uniquement les lemmes en présence ou de prendre également en compte les étiquettes sémantiques disponibles, les FP ? Le fait que deux lexies d'un même exemple soit en relation de FL, de copolysémie ou d'inclusion sémantique définitionnelle a-t-il son importance ? Il serait également important de quantifier les différences entre une telle similarité et celles basées sur les relations directement encodées par les lexicographes.

## 2.3 Analyse topologique formelle

Après une vue générale des systèmes lexicaux, nous venons de présenter en détail les éléments constitutifs du RL-fr. La nature de ses sommets et de ses arcs a été introduite, suivie d'un panorama des éléments de descriptions lexicographiques encapsulés dans ses sommets.

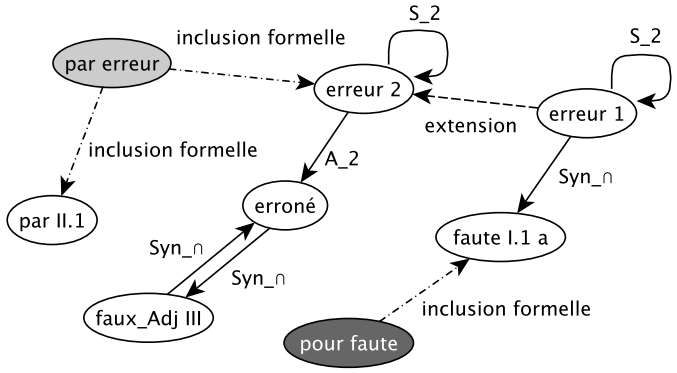



FIG. 2.24 : Représentation sagittale d'un extrait du RL-fr

La figure 2.24 propose une représentation sagittale d'un extrait du graphe lexical que nous considérons ici[19]. Rappelons que ce graphe est constitué de sommets de trois types principaux : des lexèmes, ici en blanc, des locutions, en gris clair et un petit nombre d'expressions phrasélogiques non lexicalisées, en gris foncé. Les relations qu'entretiennent ces sommets sont orientées, ce qui nous amène à les désigner comme arcs du RL-fr. Ces arcs sont également de trois types principaux : des liens de FL, des liens de copolysémie et des liens d'inclusion formelle.

Nous allons à présent nous intéresser à l'analyse topologique formelle de ce graphe lexical. Cette analyse, appelée *pedigree de graphe*, a été réalisée à l'aide du script *pedigree.py* développé par Emmanuel Navarro (Gaillard et al., 2011). Elle se déroule en deux parties. La première est consacrée aux caractéristiques formelles générales du

[19]Il est possible de considérer d'autres graphes lexicaux à partir des informations contenues dans la base de données du RL-fr. Nous pouvons ainsi imaginer un graphe dans lequel les FL seraient considérées comme des sommets, par lesquels passeraient tous les liens de FL. Les propriétés topologiques présentées dans cette section seraient alors entièrement modifées.

graphe. La seconde cherche à évaluer son appartenance à la famille des graphes petit monde. Tout comme nous l'avons fait pour la description des éléments constitutifs du graphe, nous serons ici attentive à l'évolution topologique du RL-fr dans le temps.

### 2.3.1 Caractéristiques formelles

| | |
|---|---|
| sommets | 23 599 |
| arcs | 53 081 |
| degré sortant moyen | 2,2493 |
| orienté | vrai |
| boucles | 37 |
| arcs multiples | 640 |
| arcs symétriques | 23 580 |
| sommets isolés | 1 831 |
| composantes fortement connexes | 12 987 |
| composantes faiblement connexes | 2 491 |

TAB. 2.5 : Pedigree du RL-fr du 14 août 2014 (a)

Le tableau 2.5 présente la première partie du pedigree du RL-fr, en date du 14 août 2014. Nous pouvons y voir que le RL-fr a été identifié comme étant un *multigraphe*[20], contenant des *boucles* et des *arcs multiples*. Au total, il est constitué de 23 599 sommets lexicaux et de 53 081 arcs. Rappelons que parmi ces arcs, 80% étaient alors des liens de FL, 8% des liens de copolysémie et 11% des liens d'inclusions formelles. Nous avons pu observer que la connectivité du RL-fr croît plus rapidement que sa nomenclature. La figure 2.25 montre cette évolution entre le mois de mars 2012 et le mois d'août 2014, le nombre de sommets y est représenté en pointillés, le nombre d'arc en ligne pleine. Nous pouvons y voir que le nombre d'arcs a dépassé celui de sommets dès le mois de septembre 2012, pendant la phase de constitution de la nomenclature directement induite. L'ajout de fonctionnalités permettant l'encodage de liens d'inclusion formelle et de copolysémie ne semble pas avoir eu d'impact sur ce nombre, qui poursuit une croissance régulière depuis le début de l'année 2013.

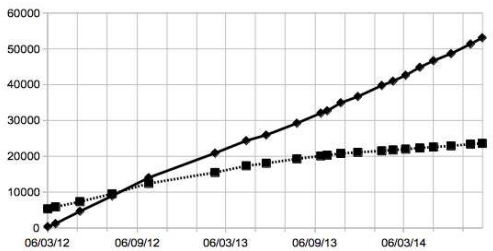



FIG. 2.25 : Évolution du nombre de sommets et d'arcs

[20] À la suite de Tabourier (2010), nous choisissons d'ignorer la distinction entre multigraphes et pseudographes. Un multigraphe est donc ici un graphe doté de boucles et d'arcs multiples.

Le *degré sortant moyen* du RL-fr mesuré est de 2,2493. Ceci signifie que, en moyenne, chaque entrée du RL-fr était alors la source d'un peu plus de deux relations. En regardant les données en détail, nous observons cependant une grande disparité de degrés entre les sommets. Ainsi, la lexie BATEAU$_{\text{N}}$ **I** [*Ce bateau navigue trois mois par an.*], pour laquelle un travail lexicographique détaillé avait été réalisé, était source de 126 arcs — trois liens de copolysémie et 123 liens de FL — tandis que 34% des sommets, encapsulant des descriptions n'ayant fait l'objet d'aucun travail particulier, n'étaient source que d'un seul arc et que près de 27% des sommets n'étaient la source d'aucun. La figure 2.26 montre l'évolution du degré sortant moyen du RL-fr entre le mois de mars 2012 et le mois d'août 2014. Nous pouvons y observer qu'il est en constante progression. Il a atteint la valeur moyenne d'un arc par sommet en août 2012, juste avant que le nombre d'arcs dépasse le nombre de sommets, et a dépassé la valeur moyenne de deux en juin 2014.

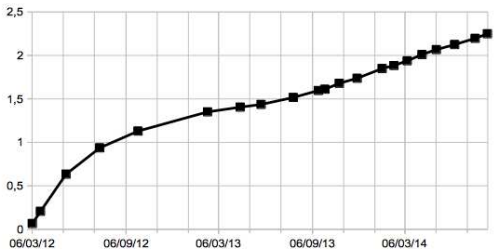



FIG. 2.26 : Évolution du degré sortant moyen

Le nombre de boucles affiché est peu élevé. Il s'agit du nombre d'arcs partant d'un sommet lexical pour revenir vers ce même sommet. Une boucle encode donc une relation particulière qu'une lexie entretient avec elle-même. De manière générale, ces boucles correspondent à des phénomènes lexicaux marginaux. Si nous nous intéressons, par exemple, à la lexie TIRAGE **II.1** [*Ils ont fait appel à une imprimerie indépendante qui réalise de très beaux tirages.*], nous pouvons voir qu'elle dispose d'une structure actancielle à trois actants, représentée à l'aide de la FP `tirage par X=1 de Y=2 en nombre Z=3`. Les lexicographes ont encodé dans sa description un lien de FL **S$_3$**, qui spécifie que son troisième actant peut être dénoté par le mot-vedette TIRAGE **II.1** lui-même. Cette dérivation sémantique nominale est notamment en œuvre dans l'exemple construit *Cet imprimeur ne réalise que de faibles tirages, mais ils sont de très bonne qualité.*

L'observation ponctuelle des boucles présentes dans le RL-fr nous a permis, à plusieurs reprises, de signaler quelques « fautes de clic » aux lexicographes. Ainsi, à la suite des observations réalisées le 14 août 2014, 13 d'entre elles se sont avérées fausses. Par exemple, le lien accidentel reliant la lexie SANITAIRE$_{\text{N}}$ **II.a** à elle-même par un lien de synonymie a été corrigé. La figure 2.27, page suivante, montre l'évolution du nombre de boucles entre le mois de mars 2012 et le mois d'août 2014. Nous pouvons y observer plusieurs vagues de corrections. La plus importante d'entre elles a eu lieu au début de l'année 2014. Elle a donné lieu à la suppression d'une dizaine de liens erronés.

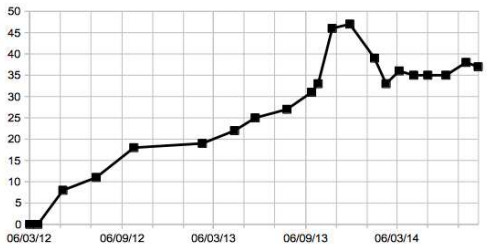



FIG. 2.27 : Évolution du nombre de boucles

Le nombre d'arcs multiples affiché est, pour sa part, plus élevé. Il rend compte de situations particulières dans lesquelles deux sommets sont reliés par plus d'un arc orientés de façon identique. Soulignons que ce nombre est calculé de manière séquentielle par le script *pedigree.py* : l'ensemble des arcs du graphe est parcouru et c'est uniquement lorsqu'un arc correspond à un couple de sommets déjà reliés par un arc de même orientation qu'il est incrémenté.

Tout comme les boucles, les arcs multiples correspondent à des phénomènes lexicaux spécifiques. Ainsi, dans le cas illustré par la figure 2.28, la lexie FRANC$_{\text{Adj}}$[1] **I.2a** [*Elle est très franche, elle ne masque pas ses sentiments.*] est la source d'un lien de copolysémie de type **extension**, qui la relie à la lexie FRANC$_{\text{Adj}}$[1] **I.2b** [*Il apporte des réponses franches et concises*]. Elle a pour FP `X=1 est franc dans Y=2`[21]. Les lexicographes ont également encodé à partir de cette lexie un lien $\mathbf{A_2}$, qui spécifie que l'adjectif typique employé pour modifier son second actant est la lexie FRANC$_{\text{Adj}}$[1] **I.2b**.

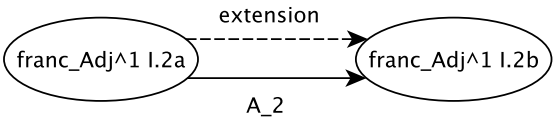



FIG. 2.28 : Exemple d'arcs multiples

L'observation des 640 arcs multiples annoncés nous a amené à comptabiliser 634 couples de lexies et 1 274 liens, parmi lesquels 151 liens de copolysémie et 200 liens d'inclusion formelle. La figure 2.29 montre l'évolution du nombre d'arcs multiples entre mars 2012 et août 2014. Nous y observons plusieurs pics de croissance. Le dernier en date, entre les mois de mai et de juin 2014 est suivi d'une diminution. Nous sommes allée regarder les données de plus près pour savoir si les liens supprimés étaient ceux qui venaient d'être ajoutés. Il n'en est rien. Nous avons en revanche pu observer que cette période correspond à un pic d'activité autour de quelques sommets fortement connectés, tels que les lexies regroupées sous les vocables SALE, VENT et APPARTEMENT.

[21] Cette FP n'était alors pas encore disponible dans la base de donnée du RL-fr, elle a donc été entrée par les lexicographes sous forme de commentaire.

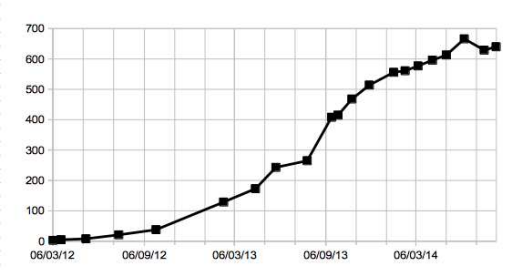



FIG. 2.29 : Évolution du nombre d'arcs multiples

Le nombre d'arcs $a \rightarrow b$ pour lesquels il existe un arc $b \rightarrow a$ affiché est beaucoup plus élevé. Nous appelons ces arcs des *arcs symétriques*. Leur observation nous a amené à comptabiliser 7 223 couples de lexies, reliés par 151 liens d'inclusion formelle, 106 de copolysémie et 14 323 de FL. La majorité de ces liens de FL correspondaient à des cas bien connus de FL disposant d'une FL inverse systématique, telles que celles discutées par Grimes (1990). Près de 13% d'entre eux étaient des liens de synonymie exacte et 11% des liens de nominalisation. Par exemple, la lexie adjectivale AUTOMOBILE$_{\text{Adj}}$ était reliée à la lexie VOITURE**I.2** à la fois par un lien de nominalisation $\mathbf{S_0}$ dont elle était la source et par un lien d'adjectivation $\mathbf{A_0}$ dont elle était la cible. La figure 2.30 montre l'évolution constante du nombre d'arcs symétriques entre mars 2012 et août 2014.

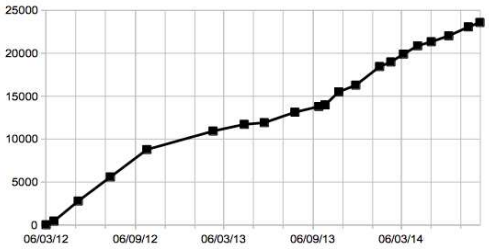
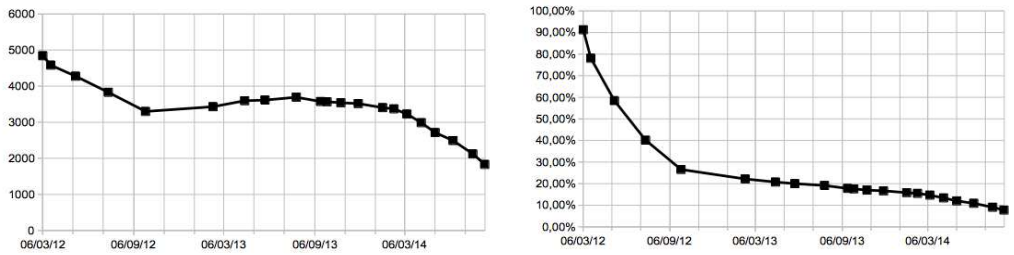



FIG. 2.30 : Évolution du nombre d'arcs symétriques

Le nombre de *sommets isolés* annoncé représente un peu moins de 8% des sommets du RL-fr. Leur observation détaillée nous amène à constater que 594 ont été créés en 2011, 703 en 2012, 516 en 2013 et seulement 18 en 2014. La moitié de ceux créés en 2011 font partie de la nomenclature d'amorçage et n'ont pas donné lieu à la création de dérivés sémantiques proches lors de l'enrichissement de la nomenclature par induction directe. Les autres ont été créés à la volée, comme cible de liens de FL. Ces liens se sont avérés incorrects et ont été supprimés. Depuis le début de l'année 2014, les lexicographes ont été invités à être vigilants afin de limiter la création de tels sommets. Lorsqu'ils créent une lexie à la volée, ils prennent le temps de lui associer au moins un lien sortant. Comme nous l'avons vu pour les arcs multiples, le travail lexicographique ne consiste pas seulement à tisser des liens, mais également à les modifier ou à les défaire. La figure 2.31 montre l'évolution de la quantité de sommets isolés entre le mois de mars 2012 et le mois d'août 2014, en nombre d'une part et en pourcentage du nombre total de sommets d'autre part. Nous pouvons y

observer que malgré une augmentation de leur nombre au cours de l'année 2013, leur proportion est restée en constante diminution. Nous constatons également que les consignes du début de l'année 2014 ont permis une importante diminution, permettant de passer sous le cap des 10% de sommets isolés au mois de juillet.




(a) Nombre (b) Proportion

FIG. 2.31 : Évolution de la quantité de sommets isolés

Les deux dernières lignes du tableau 2.5, page 74, fournissent des informations sur le nombre de *composantes connexes* du RL-fr. Il s'agit du résultat de deux décompositions consistant à partitionner le graphe lexical en sous-ensembles maximaux de sommets tous reliés entre eux. La première est une décomposition en composantes fortement connexes, pour laquelle l'orientation des arcs est prise en compte. La seconde est une décomposition en composantes faiblement connexes, pour laquelle elle ne l'est pas. La figure 2.32 illustre cette différence. Alors que le graphe qu'elle représente est constitué d'une seule composante faiblement connexe, il peut être partitionné en trois composantes fortement connexes, chacune représentée par une couleur de sommets. Comme nous pouvons le voir, un sommet unique peut former une composante. Les sommets isolés du RL-fr sont ainsi considérés comme des composantes à la fois faiblement et fortement connexes.

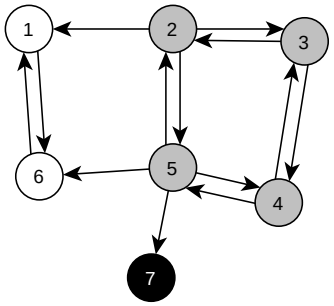



FIG. 2.32 : Composante faiblement connexe

Le modèle des systèmes lexicaux postule une organisation du lexique en un ensemble entièrement connexe, de type graphe petit monde. La possibilité de décomposer le RL-fr en composantes connexes serait donc un phénomène temporaire, lié au déroulement du travail lexicographique. Le tissage des liens à partir des lexies créées

à la volée se fait principalement sur l'axe paradigmatique. Nous pensons qu'il favorise la création de sous-ensembles correspondants à des paradigmes sémantiques[22]. Le traitement en profondeur de l'ensemble des lexies regroupées sous un même vocable, en revanche, amènerait les lexicographes à tisser des liens entre ces différents sous-ensembles, à l'aide de la copolysémie et des FL syntagmatiques.

La figure 2.33 montre l'évolution du nombre de composantes connexes entre le mois de mars 2012 et le mois d'août 2014, faiblement connexes en pointillés et fortement connexes en ligne pleine. Nous pouvons y observer une diminution conjointe depuis le dernier trimestre 2013. La période d'accélération de cette diminution pour les composantes faiblement connexes, à partir du début de l'année 2014, correspond à l'arrivée des liens de copolysémie. Nous pensons qu'il ne s'agit pas là d'une coïncidence, mais que ces liens jouent un rôle important dans la structure du lexique. Les travaux de Sigman et Cecchi (2002) exploitant la partie nominale du WordNet de Princeton vont en ce sens. Ils mettent en avant le fait que l'ajout de tels liens augmente significativement le nombre de zones denses et réduit la distance moyenne minimale à parcourir pour passer d'un sommet quelconque à n'importe quel autre. Leur approche de la polysémie est cependant différente de celle du RL-fr. Ils injectent en effet des liens symétriques entre l'ensemble des acceptions de chaque forme lexicale.

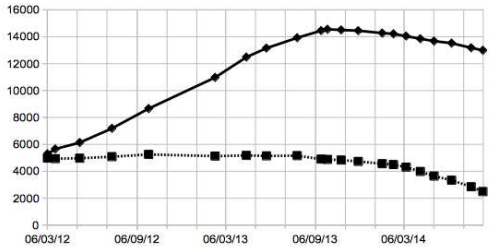



FIG. 2.33 : Évolution du nombre de composantes connexes

## 2.3.2 Graphe petit monde ?

Le tableau 2.6 présente la seconde partie du pedigree du RL-fr, en date du 14 août 2014. Elle permet de se faire une idée sur l'organisation des sommets et des arcs à l'intérieur du RL-fr. Comme nous l'avons vu, les systèmes lexicaux sont théoriquement apparentés aux graphes de données réelles observés dans de nombreux domaines, appelés *graphes petit monde*. De tels graphes ont été identifiés par Watts et Strogatz (1998) et notamment étudiés par Newman (2003), Delahaye (2003), Gaume (2004), Tabourier (2010) et Navarro (2013). Ils se distinguent par la concomitance de trois caractéristiques.

1. une faible densité, c.-à-d. un petit nombre d'arcs relativement au nombre de sommets ;

[22]Cette hypothèse a fait l'objet d'une expérience, présentée dans le chapitre 4.

| | |
|---|---|
| coefficient d'agrégation | 0,276 |
| Plus grande composante fortement connexe | |
| sommets | 5 540 |
| arcs | 19 560 |
| longueur moyenne des plus courts chemins | 13,0255 |
| Distribution des degrés sortants | |
| a | -2,2401 |
| $r^2$ | 0,9550 |
| Distribution des degrés entrants | |
| a | -2,4821 |
| $r^2$ | 0,9286 |

TAB. 2.6 : Pedigree du RL-fr du 14 août 2014 (b)

2. un coefficient d'agrégation élevé, c.-à-d. une forte probabilité que deux sommets voisins d'un même sommet soient eux-mêmes voisins ;

3. une faible moyenne des plus courts chemins entre deux sommets quelconques du graphe.

Pour déterminer la *densité* d'un graphe, il faut s'intéresser aux nombres d'arcs ($m$) et de sommets ($n$) qui le constituent. Si le RL-fr ne comportait ni boucle, ni arc multiple, il serait alors considéré comme un graphe simple. Dans un graphe simple, la densité maximale correspond au cas où chaque sommet est source d'un arc vers chacun des autres sommets, soit $n \times (n-1)$ arcs. Pour un graphe de 23 599 sommets, cela correspond donc à environ 556 x $10^6$ arcs. Nous pouvons donc affirmer que le RL-fr, avec 53 081 arcs, avait une faible densité le 14 août 2014. Gaume (2004) va plus loin, en affirmant que le nombre d'arcs observés dans un graphe petit monde est généralement inférieur à $nlog(n)$. Là encore, la densité du RL-fr était conforme à celle attendue. Nous avons vu que la connectivité du RL-fr croît plus rapidement que sa nomenclature. La figure 2.34 présente l'évolution du rapport entre $nlog(n)$ et $m$ entre le mois de mai 2012 et le mois d'août 2014[23]. Nous pouvons y observer que le nombre d'arcs du RL-fr se rapproche progressivement, mais lentement, de cette valeur limite.

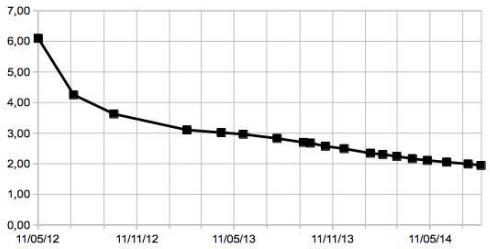



FIG. 2.34 : Évolution de la densité

[23]Au début du mois de mars 2012, $nlog(n)$ était 56 fois supérieur au nombre d'arcs et 18 fois supérieur à la fin du même mois. Ces valeurs ont été exclues de la figure par soucis de lisibilité.

Le *coefficient d'agrégation*, en anglais *clustering coefficient*, permet d'estimer dans quelle mesure deux sommets d'un graphe reliés à un même sommet sont également directement reliés entre eux. Si ce coefficient est élevé, cela signifie que le graphe est robuste et qu'il est possible de supprimer des sommets sans que ni la distance maximale entre deux sommets quelconque, ni la moyenne des distances minimales entre l'ensemble des sommets ne varient significativement. Cela signifie également que le graphe est composé de zones denses.Selon Newman (2003), le coefficient d'agrégation d'un graphe doit être considéré relativement à celui d'un graphe aléatoire classique de même densité. Delahaye (2003) estime la valeur d'un tel coefficient à $2m/n^2$. Un graphe aléatoire composé de 23 599 sommets et 53 081 arcs aurait un coefficient d'agrégation proche de 0,00018. Le 14 août 2014, le RL-fr présentait donc un coefficient élevé.

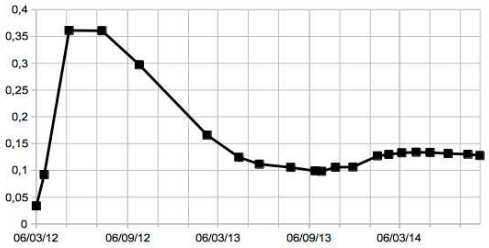



FIG. 2.35 : Évolution du coefficient d'agrégation

La figure 2.35 montre son évolution entre mars 2012 et août 2014. Nous pouvons y observer qu'elle est irrégulière. Alors qu'elle débute par une forte augmentation, elle chute à partir du mois de septembre 2012, puis recommence à progresser à la fin de l'année 2013 pour se stabiliser aux alentours de 0,13. Le coefficient d'agrégation du RL-fr demeure néanmoins toujours largement supérieur à celui de graphes aléatoires de même densité.

La *moyenne des plus courts chemins* permet de se faire une idée sur l'organisation des agrégats au sein du graphe. Une faible moyenne signifie qu'il est possible de passer rapidement d'un sommet du graphe à n'importe quel autre. Elle est mesurée selon la formule suivante[24], où $l_{u,v}$ représente le plus petit nombre d'arcs qu'il est nécessaire d'emprunter pour se rendre du sommet $u$ au sommet $v$, en suivant l'orientation des arcs :

$$\ell = \frac{1}{n(n-1)} \sum_{u,v \in V} l_{u,v}$$

Selon Watts et Strogatz (1998), la moyenne des plus courts chemin d'un graphe petit monde est comparable à celle d'un graphe aléatoire de même taille et de même degré moyen. Newman (2001) reprend cette idée en la nuançant, même si

[24]Nous empruntons cette formule à Navarro (2013). Elle est légérement différente de celle proposée par Newman (2003), qui inclue également la distance minimale entre chaque sommet et lui-même, valant invariablement zéro.

ces moyennes sont proches, elles sont sensiblement différentes. Selon lui, la moyenne des plus courts chemins dans un graphe aléatoire peut être estimée à l'aide de la formule suivante, où $n$ représente le nombre de sommets du graphe et $z$ son degré moyen :

$$\ell = log\, n / log\, z$$

Dans la majorité des graphes de collaborations scientifiques qu'il analyse, la moyenne des plus courts chemins est sensiblement supérieure à cette valeur.

Comme le souligne Newman (2003), la mesure de la moyenne des plus court chemin est problématique dans le cas de graphes non connexes, tels que le RL-fr. Plusieurs alternatives sont possibles pour pallier cette difficulté. L'une d'entre elles, exploitée par Newman (2003), consiste à effectuer la moyenne de l'ensemble des plus courts chemins qu'il est possible de calculer et de faire abstraction des couples de sommets non connectés. Une seconde, disponible à travers la fonctionnalité *average_path_length* de la librairie python *igraph*[25], consiste à calculer la moyenne des longueurs moyennes des plus courts chemins de l'ensemble des composantes du graphe. Pour le RL-fr du 14 août 2014, cette méthode retourne une valeur de 13,3097. Une troisème alternative, à laquelle le script *pedigree.py* nous donne accès, consiste à effectuer cette mesure sur la plus grande partie connexe du graphe analysé[26]. Comme nous l'avons vu au cours de la section 2.2.2, l'orientation des arcs joue un rôle important dans la structuration du RL-fr. Pour cette raison, nous avons choisi de nous intéresser à la plus grande composante fortement connexe, plutôt qu'à sa plus grande composante faiblement connexe. Le 14 août 2014, la valeur obtenue était proche du résultat de la méthode précédente[27] : 13,0255. Dans les deux cas, la valeur considérée est sensiblement supérieure à celle d'un graphe alétoire de même taille et de même degré moyen : $log(23599)/log(2,249) \simeq 12,4234$. Elle est donc conforme aux observations de Newman (2001).

Newman (2003) définit l'« effet petit monde » en terme d'évolution de la moyenne des plus courts chemin au cours de l'extension d'un graphe dont le degré moyen reste constant.

> If the number of vertices within a distance $r$ of a typical central vertex grows exponentially with $r$ [...] then the value of $\ell$ will increase as $log\, n$. In recent years the term "small-world effect" has thus taken on a more precise meaning : networks are said to show the small-world effect if the value of $\ell$ scales logarithmically or slower with network size for fixed mean degree.
>
> Newman (2003)

Nous aurions souhaité comparer l'évolution de la moyenne des plus courts chemins du RL-fr à cette estimation. Cependant, comme nous l'avons vu dans la section

[25] http ://igraph.org/python/

[26] Nous aurions également pu envisager d'ajouter manuellement un certain nombre de liens afin de rendre le RL-fr entièrement connexe. Une réflexion aurait alors été nécessaire, en concertation avec les lexicographes, afin d'effectuer des connexions pertinentes.

[27] La moyenne des plus courts chemin de la plus grande composante faiblement connexe était pour sa part de 13,3102.

2.3.1, le degré moyen du RL-fr est pour l'heure en constante progression.

Le tableau 2.7 montre l'évolution de l'étendue du RL-fr et de la moyenne de ses plus courts chemins, ici mesurée à l'aide de la fonctionnalité *average_path_length* de la librairie python *igraph*, entre le 30 juillet 2013 et le 14 août 2014. Nous pouvons y observer que, tandis que le nombre de sommets de cette composante a augmenté de plus de 20% en un an, le nombre d'arcs de plus de 80% et le degré sortant moyen de ses sommets de près de 50%, la moyenne des plus courts chemins a pour sa part diminué de près d'un quart. Ces observations nous permettent uniquement de constater que la moyenne des plus courts chemins du RL-fr diminue tandis que le réseau s'étend.

| | 30/07/13 | 14/08/14 | évolution |
|---|---|---|---|
| sommets | 19 280 | 23 599 | +22% |
| arcs | 29 232 | 53 081 | +82% |
| degré moyen sortant | 1,5162 | 2,2493 | +48% |
| longueur moyenne | 17,4137 | 13,3097 | −24% |

TAB. 2.7 : Évolution de la longueur moyenne des plus courts chemins

Une quatrième caractéristique du RL-fr est disponible dans la seconde partie de son pedigree. Elle nous éclaire également sur l'organisation des agrégats au sein du graphe. Plus précisément, elle concerne la distribution des probabilités du nombre d'arcs associés à un sommet, ou *distribution des degrés* sortants et entrants. Elle est mesurée en associant à chaque entier $k$ le nombre de sommets du graphe ayant un degré $k$. Cette distribution suit une loi de puissance si elle est proportionnelle à $k^{-\gamma}$ pour un certain réel $\gamma$, appelé *exposant de la loi*. Les graphes petit monde qui disposent d'une telle caractéristique sont considérés comme des graphes petit monde *hiérarchiques*[28] (Ravasz et Barabási, 2003). Le 14 août 2014, la distribution des degrés sortants du RL-fr était fortement corrélée (0,9550) à une loi de puissance de coefficient -2,2401. La distribution des degrés entrants était elle aussi fortement corrélée (0,9286) a une loi de puissance de coefficient -2,4821. Cela signifie que, conformément à nos observations précédentes, un petit nombre de sommets étaient fortement connectés, tandis qu'un grand nombre d'entre eux l'étaient peu. L'observation de la distribution des degrés entrants du RL-fr nous amène à constater que les lexies qui étaient alors la cibles de nombreux liens étaient des lexies carrefours jouant un rôle central dans l'organisation du lexique, tandis que les lexies qui l'étaient peu étaient des lexies rares. Ainsi, la lexie FAIRE **II.1** [*Il fait du ping-pong.*] était cible de 155 liens, dont 119 liens de FL et 36 liens d'inclusion formelle. À l'opposé, la lexie MONOGRAMME [*Cette assiette est signée PAK, monogramme de Pieter Adriaensz Kocks.*] n'était la cible que d'un seul lien (de FL). Cette propriété se vérifie pour l'ensemble des versions du RL-fr, de mars 2012 à août 2014.

[28]Notons que Bollobás et Riordan (2004) ont montré que le rythme de croissance de la moyenne des plus courts chemins des graphes petit monde hiérarchiques n'excède pas $\log n / \log \log n$.

## Conclusions

Nous avons présenté, tout au long de ce chapitre, les propriétés du Réseau Lexical du Français. Après avoir introduit le modèle des systèmes lexicaux, dont il est une instance, nous avons présenté en détail l'ensemble de ses éléments constitutifs et une partie de ses caractéristiques formelles. Comme nous l'avons vu, cette ressource est en cours de développement. En date du 14 août 2014, il s'agissait d'un multigraphe orienté constitué de 23 599 sommets et de 53 081 arcs. Seul un noyau de 263 sommets disposait d'une description arrêtée — la plus complète possible au vu des avancées de l'éditeur lexicographique Dicet — 8 579 étaient en cours de description et 14 757 étaient à peine créés. Plus de 80% des arcs correspondaient à des liens sémantico-syntaxiques de fonctions lexicales Sens-Texte, 8,5% à des liens sémantiques de copolysémie et 11% à des liens asémantiques d'inclusion formelle. Nous avons pu constater à plusieurs reprises que ces liens étaient non seulement tissés, mais également détissés par les lexicographes à partir des lexies en cours de description. Nous avons également observé que les éléments des descriptions lexicographiques encapsulées dans les sommets ne bénéficiaient pas tous de la même couverture. L'appartenance à un vocable était encodée pour chaque sommet et les caractéristiques grammaticales l'étaient de façon quasi-systématique, avec un indice de confiance maximal dans plus de 99% des cas. Des exemples d'emploi étaient associés à 80% des sommets, mais 76% d'entre eux n'en disposaient que d'un seul. Les étiquettes sémantiques couvraient 70% de la nomenclature, mais seulement 40% d'entre elles étaient encodées avec un indice de confiance maximal. Enfin, plus de la moitié des sommets disposaient d'une FP et un cinquième disposaient d'informations morphologiques flexionnelles.

Une analyse des propriétés formelles de ce multigraphe nous a permis de mettre en avant que son organisation était proche de celle d'un graphe petit monde hiérarchique. Il était peu dense, mais organisé en sous-ensembles de sommets fortement interconnectés. Un grand nombre de ses sommets avaient un faible degré et un petit nombre de sommets un degré élevé. Il ne s'agissait pas d'un graphe connexe. La moyenne de ses plus courts chemins était proche de celle d'un graphe aléatoire de même taille et de même degré moyen. Nous avons pu observer qu'elle diminue au cours du temps, mais cette évolution étant concomitante à une augmentation du degré moyen, nous n'avons pas pu évaluer si elle était conforme à celle des graphes petits monde généralement observés.

Cette présentation nous a permis de circonscrire les objets mis en relation dans la ressource que nous souhaitons explorer par raisonnement analogique. Comme nous l'avons vu, nous pensons que les relations encodées entre ces objets peuvent être exploitées pour mettre au point des mesures de similarité qui guideront cette exploration. En revanche, les éléments des descriptions lexicographiques ne nous semblent pas tous a priori pertinents pour établir des mesures de similarité d'Attributs. Lors de l'exposé de nos expériences qui va suivre, nous préciserons chaque fois que nécessaire les éléments pris en compte et leur disponibilité dans la version du RL-fr exploitée. Les relations en jeu dans les résultats obtenus seront explicitées au fur et à mesure. La majorité des FL évoquées sont également disponibles en annexe.

# Chapitre 3

# Expérience initiale

## Sommaire

# Introduction

Ce chapitre présente notre première expérience d'automatisation du raisonnement analogique sur des données lexicographiques. Étant donné un ensemble défini de descriptions de lexies, nous avons cherché à détecter automatiquement des proportions analogiques correspondant à l'analogie suivante :

la description de la LEXIE **A** est à la description de la LEXIE **B**
ce que
la description de la LEXIE **C** est à la description de la LEXIE **D**

L'objectif de cette expérience était de confronter nos lectures sur le raisonnement analogique au RL-fr, d'affiner notre compréhension de celles-ci et de tester si les hypothèses qui en découlaient pouvaient trouver un écho dans la mise en œuvre d'un prototype.

Au moment où cette expérience a été réalisée, la description des lexies n'était pas encore aussi détaillée que ce qui a été présenté au chapitre 2. Nous disposions seulement de trois types d'information[1] : l'appartenance d'une lexie à un vocable, ses CG et les liens de FL la reliant aux autres lexies.

Ce chapitre présente les aspects théoriques qui ont guidé nos choix méthodologiques et les hypothèses que nous avons souhaité tester. Cette présentation est faite au regard de notre état de compréhension avant de mener l'expérience. Certaines considérations s'avéreront donc erronées et divergent de ce que nous discutons dans les autres chapitres. Ce chapitre détaille ensuite les données sélectionnées, le format de représentation de ces données et la mise en œuvre technique de l'expérience. Enfin, une analyse des résultats est proposée.

Cette expérience pose les bases d'une méthode de détection de similarités et de relations analogiques entre lexies qui sera affinée par la suite.

## 3.1 Hypothèses

Comme nous l'avons vu dans le chapitre 1, les travaux d'Yves Lepage ont été un point d'entrée précieux pour nous approprier la notion d'*analogie*. À sa suite (Lepage, 2001), nous abordons cette notion comme « une égalité de rapports ».

> Une analogie est une égalité de rapports énoncée par « $A$ est à $B$ ce que $C$ est à $D$ et notée par A : B = C : D. ».
>
> Lepage (2001, p.1)

Dans son mémoire d'HDR (Lepage, 2003), qui contient une présentation historique détaillée de cette notion, Lepage relève l'importance de la *similarité* et de la *contiguïté* — ici employées à la suite de Jakobson (1963), l'une étant le fondement de la sélection, l'autre celui de la combinaison — dans la notion d'*analogie*.

[1] Lorsque cette expérience a été réalisée, nous ne considérions pas les FL comme des relations, mais comme des éléments de descriptions de lexies.

> (...) l'analogie, en tant que position intermédiaire, doit nécessairement mettre en jeu à la fois la similarité et la contiguïté.
>
> Lepage (2003, p.99)

Nous avons choisi de nous intéresser, dans un premier temps, à la mise au point d'une mesure de similarité et de laisser de côté la question de la contiguïté.

### 3.1.1 Mesure de similarité

Nous avons fait l'hypothèse que la similarité entre lexies pouvait être mesurée à l'aide d'une mesure de *distance d'édition*.

Pour une première expérience, il nous a semblé intéressant de mettre à plat toute la complexité de notre structure de données et de considérer les descriptions de lexies comme des chaînes d'informations assimilables à des chaînes de caractères. Pour une telle approche, nous disposions d'un ensemble de descriptions, élaborées à l'aide d'un vocabulaire fini non ordonné[2] (ici un ensemble fini de 9 CG + un ensemble fini de 100 FL présentés dans les tableaux de la section 3.2), dont nous avions alors admis un nombre infini de combinaisons possibles. L'algorithme de Wagner et Fischer (1974), implémentant la mesure de distance d'édition, était alors applicable à nos données.

#### 3.1.1.1 Distance d'édition

La distance d'édition, autrement connue sous le nom *de distance de Levenshtein*[3], consiste à calculer le nombre d'opérations minimales nécessaires pour obtenir une chaîne de caractères B à partir d'une chaîne de caractères A. Nous allons à présent en dérouler le fonctionnement. Le calcul de distance d'édition autorise trois types d'opérations sur la chaîne A :

- l'ajout d'un caractère nouveau (qui est présent en B),
- la suppression d'un caractère de A (qui est absent de B),
- la substitution d'un caractère de A par un caractère de B.

Chacune de ces opérations a un coût égal à 1. Conserver un caractère de A (car présent dans B), autrement dit le substituer à lui-même, revient à n'effectuer aucune opération et coûte donc 0. Le coût total des opérations est équivalent que nous souhaitions obtenir B à partir de A ou A à partir de B.

Soit A la chaîne **tua** et B la chaîne **use**, obtenir la chaîne B à partir de la chaîne A nécessite les quatre opérations suivantes :

1. suppression de **t** : 1
2. substitution de **u** par **u** : 0
3. ajout de **s** : 1

[2] Autrement dit, nous estimions alors qu'aucun élément du vocabulaire n'avait plus d'importance qu'un autre.

[3] Du nom de Vladimir Levenshtein, qui l'a définie en 1965 (voir Levenshtein, 1966).

4. substitution de **a** par **e** : 1

La distance d'édition entre A et B est donc égale à 3.

Si nous mesurions de la sorte la distance entre descriptions de lexies, nous aurions obtenu une valeur d'autant plus élevée que les différences entre elles étaient nombreuses.

Une difficulté se serait alors présentée. La distance d'édition n'est pas une valeur réelle comprise entre 0 et 1, ni un pourcentage. Elle dépend donc en partie de la longueur des descriptions comparées. Telle quelle, elle ne nous aurait pas permis d'identifier directement des couples de descriptions de lexies également similaires.

#### 3.1.1.2 De la similitude à la similarité

Pour pallier cette difficulté et obtenir un degré de similarité entre descriptions de lexies, nous nous sommes appuyée sur une variante de la distance d'édition, la *distance d'édition canonique* proposée par Lepage (2003, p.102).

> Il existe une distance d'édition bien particulière qui entretient un rapport privilégié avec la similitude. Il s'agit de la distance équipée seulement de deux opérations d'insertion et de suppression.
>
> Lepage (2003, p.102)

Dans cette variante, la substitution n'est plus disponible comme opération de coût 1. Pour substituer un caractère $b$ de la chaîne B à un caractère $a$ de la chaîne A, il faut donc supprimer $a$, puis ajouter $b$, ce qui a un coût de 2.

Cette définition permet à Lepage de proposer une mise en relation directe de la *dissimilarité* de deux chaînes et de leur *similarité*.

La distance d'édition canonique, notée $\delta(A, B)$, correspond aux coûts cumulés des insertions et des suppressions de caractères nécessaires pour passer de la chaîne A à la chaîne B.

Divisée par le cumul des longueurs de A et B, elle permet d'obtenir une mesure de *dissimilarité*.

La *similitude*, notée $\sigma(A, B)$, correspond à la longueur du plus long sous-mot commun aux deux chaînes, c'est-à-dire à l'addition des longueurs de toutes leurs sous-chaînes communes. La *similarité* est obtenue en ramenant la similitude à une valeur réelle comprise entre 0 et 1.

En appliquant ces nouvelles mesures aux chaînes A **tua** et B **use** de l'exemple précédent, nous obtenons les résultats présentés dans le tableau 3.1.

| Nom de la mesure | Notation | Applications aux chaînes | Résultat |
|---|---|---|---|
| Distance d'édition canonique | $\delta(A, B)$ | $2\,suppressions\,[\mathbf{t}, \mathbf{a}]+$ $2\,insertions\,[\mathbf{s}, \mathbf{e}]$ | 4 |
| Dissimilarité | | $\frac{\delta(A,B)}{\|A\|+\|B\|}$ | $\frac{4}{6} \simeq 0,66$ |
| Similitude | $\sigma(A, B)$ | $1\,caractère\,commun\,[\mathbf{u}]$ | 1 |
| Similarité | | $\frac{2\times\sigma(A,B)}{\|A\|+\|B\|}$ | $\frac{2}{6} \simeq 0,33$ |

TAB. 3.1 : Application de la mesure de similarité aux chaînes A **tua** et B **use**

Nous vérifions alors bien la proposition de Lepage (2003) selon laquelle :

$$\forall(A, B) \in (\mathcal{V}^*)^2, \delta(A, B) = |A| + |B| - (2 \times \sigma(A, B))$$

$$\delta(tua, use) = 4 = |A| + |B| - (2 \times \sigma(tua, use)) = 3 + 3 - (2 \times 1)$$

Ainsi que la complémentarité entre similarité et dissimilarité :

$$similarité(tua, use) = \tfrac{2}{6} = 1 - dissimilarité(tua, use) = 1 - \tfrac{4}{6}.$$

À cette étape, nous disposions de deux mesures directement utilisables : la dissimilarité et la similarité. Nous pouvions alors envisager de distinguer des couples de descriptions de lexies plus ou moins similaires et d'effectuer des regroupements entre ces couples.

### 3.1.2 Similarité d'Attributs et similarité de Relations

Pour affiner notre compréhension du raisonnement analogique, nous nous sommes intéressée aux travaux en sciences cognitives, notamment à ceux de Drede Gentner.

Comme nous l'avons vu dans la section 1.3 du premier chapitre, Medin, Goldstone et Gentner (1990) établissent l'existence d'une préférence cognitive à établir les jugements de similarité à partir d'assortiments de Relations, tandis que l'absence d'assortiments d'Attributs serait privilégiée pour les jugements de différenciation.

> Logically, our pattern of results could be produced by either or both of the following: (a) attributional matches are less important than relational matches in similarity judgment, and (b) attributionnal mismatches are more important than relational mismatches for dissimilarity judgments.
>
> Medin et al. (1990, p.68)

Il nous a semblé que, en admettant que chaque description de lexie correspondait à un stimulus, nous pouvions considérer les mesures de dissimilarité et de similarité présentées en 3.1.1 comme des comparaisons d'Attributs.

Les Relations susceptibles de se dégager étaient alors telles que « les FL en jeu dans la description de cette lexie sont toutes des FL verbales » ou encore « les CG

sont au nombre de 2 ».

L'approche choisie ne nous permettait pas d'établir directement des mesures de similarité de Relations. Nous avons tout de même souhaité tester l'hypothèse selon laquelle un degré de similarité d'Attributs (désormais $sim_a$) élevé entre descriptions de lexies était un indice de Relations pertinentes entre leurs Attributs et permettait donc de simuler des jugements de similarité, tout autant que de différenciation.

Nous distinguions alors trois ordres de Relations : des Relations internes propres à chaque description de lexie, externes communes aux deux descriptions dont le degré de $sim_a$ est élevé ou externes communes à toutes les descriptions partageant un tel degré de $sim_a$ avec au moins une autre description.

### 3.1.3 Relations analogiques

À ce stade d'élaboration du prototype, nos choix concernant l'automatisation d'un jugement de similarité entre descriptions de lexies étaient fixés. L'objectif de l'expérience n'était cependant pas la détection de tels jugements, mais celle de proportions analogiques, mettant en œuvre non pas deux, mais quatre éléments, ici des descriptions de lexies.

#### 3.1.3.1 Proportions analogiques

Rappelons-le formellement, l'objectif de cette expérience était la détection de proportions analogiques correspondant à l'analogie suivante :

la description de la LEXIE **A** est à la description de la LEXIE **B**
ce que
la description de la LEXIE **C** est à la description de la LEXIE **D**

Étant donné la richesse et la rigueur des descriptions des lexies du RL-fr, nous pensions possible de détecter automatiquement des proportions rendant compte de relations analogiques entre lexies. Une telle relation existe notamment entre les paires de lexies CONDUIRE et VOITURE **1** d'une part et PILOTER **I** et AVION d'autre part, dont il est possible de dire :

CONDUIRE est à VOITURE **1**
ce que
PILOTER **I** est à AVION

La mesure de $sim_a$ proposée en 3.1.1 ne semblait pas directement exploitable pour obtenir automatiquement une telle proportion. Nous pouvions en revanche nous attendre à ce qu'elle attribue un degré de $sim_a$ élevé au couple de descriptions des lexies CONDUIRE et PILOTER **I** tout autant qu'au couple de descriptions des lexies VOITURE **1** et AVION.

Nous nous sommes alors intéressée aux propriétés fondamentales des proportions analogiques. Stroppa (2005) rappelle ses propriétés et l'ensemble d'expressions équivalentes défini par Lepage (2003). Soit la proportion analogique $A : B :: C : D$ ;

nous listons ici l'ensemble de ces expressions, associées aux conditions dont elles découlent, sans détailler davantage :

*Symétrie de la conformité* : $C : D :: A : B$

*Permutation des moyens* : $A : C :: B : D$

*Inversion des rapports* : $B : A :: D : C$

*Permutation des extrêmes* : $D : B :: C : A$

*Symétrie de lecture* : $D : C :: B : A$, $B : D :: A : C$, $C : A :: D : B$

Parmi cet ensemble d'expressions équivalentes[4], la *permutation des moyens* nous a particulièrement intéressée. En effet, elle permet de reconstruire la proportion analogique CONDUIRE : VOITURE**1** :: PILOTER**I** :AVION à partir des deux couples de lexies de fort degré de $sim_a$ (CONDUIRE, PILOTER**I**) et (VOITURE**1**, AVION).

Nous pouvions alors estimer que les proportions analogiques détectées à partir de couples de descriptions de $sim_a$ élevées étaient valides. En effet, s'il semble moins naturel d'énoncer « CONDUIRE est à PILOTER**I** ce que VOITURE**1** est à AVION », la proportion analogique rendant compte de cette relation contient toutes les informations nécessaires à l'énonciation de son équivalent « CONDUIRE est à VOITURE**1** ce que PILOTER**I** est à AVION ».

#### 3.1.3.2 Synonymie et analogie

La synonymie est un cas particulier de similarité forte entre lexies et par conséquent entre descriptions de lexies[5]. À partir de notre méthode de comparaison, nous pouvions prévoir que les descriptions de lexies synonymes aient une $sim_a$ très élevée. En plus de partager un certain nombre de CG et de FL en jeu dans leur description, nous savions qu'elles partageraient un nombre important de valeurs d'application de FL. Nous pouvions nous attendre, par exemple, à ce que les lexies VOITURE**1** et AUTO**I** aient toutes deux CONDUIRE comme valeur d'application de la FL $\mathbf{Real_1}$.

Nous souhaitions toutefois éviter la construction automatique de proportions telles que « CONDUIRE est à VOITURE**1** ce que CONDUIRE est à AUTO**I** ».

Turney (2006), qui s'intéresse à l'analogie entre mots, associe l'analogie aux similarités de Relations, tandis que les similarités d'Attributs seraient à rapprocher de la synonymie.

> When two words have high degree of attributional similarity, we call them *synonyms*. When two word *pairs* have high degree of relational similarity, we say they are *analogous*.

[4]Notez que, dans l'ensemble des expressions équivalentes proposées, à aucun moment $A$ et $D$ ne sont mis en relation directe, pas plus que $B$ et $C$.

[5]À l'époque où nous avons mené cette expérience, deux lexies en lien de synonymie exacte étaient toutes deux entièrement décrites par les lexicographes. À présent, dans un certain nombre de cas, seule l'une d'entre elles l'est, la description de la seconde ne comporte que les informations qui lui sont propres. C'est le cas notamment des lexies W.C.**a** et CABINET**III.a**. Seule la lexie W.C.**a** est entièrement décrite. La description de la lexie CABINET**III.a** comporte des CG, une étiquette sémantique, une FP, la FL **Syn** et ses cibles, ainsi que des exemples. Les autres fonctions lexicales sont à consulter dans la description de W.C.**a**.

Turney (2006, p.379)

Reprenons l'exemple précédent. Si nous nous intéressons aux descriptions des lexies VOITURE **1** et CONDUIRE d'une part et AVION et PILOTER **I** d'autre part, nous constatons que ce n'est plus précisément l'Attribut « valeur d'application de la FL **Real$_1$** » qui est important, mais la Relation « La seconde lexie de la paire apparaît dans la description de la première en tant que valeur d'application de la FL **Real$_1$** »[6].

Ne disposant pas de mesure de similarité de Relations, nous avions fait le choix de privilégier la recherche de paires de lexies appartenant à un même vocable. Toute synonymie étant exclue dans ce cadre, la mesure de $sim_a$ choisie pouvait être utilisée sans risque de confusion entre descriptions de lexies synonymes et paire de descriptions de lexies entrant en relation d'analogie.

Comme nous l'avons vu au chapitre 2, nous savions par ailleurs que les lexies regroupées au sein d'un même vocable entretiennent des relations sémantiques[7]. Nous avions alors fait l'hypothèse que l'appartenance à un même vocable était un Attribut fort, favorisant l'apparition de Relations pertinentes entre les autres Attributsdes descriptions de lexies comparées, que ces Relations soient simplement internes aux lexies ou communes à l'intérieur du vocable.

Ainsi, nous nous trouvions en mesure de tester si les descriptions de lexies d'un même vocable ayant un certain degré de $sim_a$ entraient en relation analogique avec les descriptions de lexies de tout autre vocable ayant entre elles une $sim_a$ analogue.

À partir de ces deux hypothèses, nous avons concentré notre attention sur les couples de descriptions de lexies de même vocable ayant un degré de $sim_a$ supérieur à 0,75 d'une part, et compris entre 0,5 et 0,749 d'autre part[8].

#### 3.1.3.3 Organisation des proportions analogiques

À ce stade, nous disposions d'une mesure de $sim_a$ et savions sur quels couples de $sim_a$ élevée concentrer notre attention. Nous n'avions cependant pas encore résolu la question de l'organisation des éléments des couples dans les proportions analogiques.

La mesure pour laquelle nous avons opté n'est pas une mesure orientée :

$$\forall(A, B) sim_a(A, B) = sim_a(B, A).$$

Ainsi, en présence de deux couples de descriptions de lexies telles que celles des deux acceptions[9] du vocable ACCOMPAGNEMENT de $sim_a \simeq 0,71$ et celles des

[6] Le modèle de représentation utilisé lors de cette première expérience ne permet pas de rendre compte d'une telle relation.

[7] Voir Polguère (2008, p.155-157) pour une présentation des notions d'homonymie et de polysémie et la question du regroupement des lexies sous un même vocable.

[8] Ces seuils ont été choisis de façon arbitraire à la suite d'une observation rapide des résultats.

[9] Les deux acceptions du vocable ACCOMPAGNEMENT dont il est question ici sont les lexies ACCOMPAGNEMENT **I** [*La mairie assure l'accompagnement des enfants jusqu'au centre de loisir.*] et ACCOMPAGNEMENT **II** [*Est-ce que tu penses qu'il a besoin d'un accompagnement pour arrêter de fumer ?*] Les deux acceptions du vocable DIAGNOSTIC sont les lexies DIAGNOSTIC **I** [*Le diagnostique est tombé : Max est atteint d'Ebola.*] et DIAGNOSTIC **II** [*Nous devrions recevoir le diagnostic financier dans la journée*].

deux acceptions du vocable DIAGNOSTIC de $sim_a \simeq 0,67$, nous aurions pu établir indifféremment les deux proportions analogiques suivantes, ainsi que l'ensemble des expressions équivalentes :

description de ACCOMPAGNEMENT II : description de ACCOMPAGNEMENT I
::
description de DIAGNOSTIC II : description de DIAGNOSTIC I

description de ACCOMPAGNEMENT II : description de ACCOMPAGNEMENT I
::
description de DIAGNOSTIC I : description de DIAGNOSTIC II

Nous savions cependant, à la suite de Mel'čuk, Clas et Polguère (1995), que les lexies sont organisées à l'intérieur des vocables, qui contiennent tous une unité lexicale de base.

> La **lexie de base** d'un vocable est une lexie L telle que les autres lexies du vocable font directement ou indirectement référence à L alors que L ne fait aucune référence aux autres lexies du vocable.
>
> Mel'čuk et al. (1995, p.159)

Nous avons alors estimé qu'il était utile d'en tenir compte dans l'organisation des proportions analogiques. Les proportions analogiques permettant de faire émerger des règles pertinentes seraient celles qui mettaient en relation des descriptions d'unités lexicales de base (désormais DULB) et des descriptions d'unités lexicales secondaires (désormais DULS) de la manière suivante :

$$DULS_{voc1} : DULB_{voc1} :: DULS_{voc2} : DULB_{voc2}$$

En d'autres termes, nos hypothèses nous amenaient à considérer l'existence de relations analogiques telles que « La description de la lexie ACCOMPAGNEMENT II est à la description de la lexie ACCOMPAGNEMENT I ce que la description de la lexie DIAGNOSTIC II est à la description de la lexie DIAGNOSTIC I ».

#### 3.1.3.4 Degré d'analogicité

Une fois l'ensemble des hypothèses précédentes établies, nous avons souhaité nous munir d'un nouvel indice nous permettant d'évaluer la qualité des proportions analogiques détectées.

Nous nous sommes alors intéressée aux travaux de Turney (2006). À la suite de Gentner (1983) et Medin et al. (1990), Turney (2006) distingue des analogies *proches* et des analogies *lointaines*. Comme nous l'avons vu dans le premier chapitre, il s'appuie pour cela sur le degré de $sim_a$ entre les membres $A$ et $C$ d'une part et $B$ et $D$ d'autre part, pour déterminer la proximité de l'analogie correspondant à la proportion $A : B :: C : D$ et lui attribuer un score.

> In analogy $A : B :: C : D$, where there is a high degree of relational similarity between $A : B$ and $C : D$, if there is also a high degree of attributional similarity between $A$ and $C$, and between $B$ and $D$, then $A : B :: C : D$ is a near analogy; otherwise, it is a far analogy.
>
> Turney (2006, p.383)

Nous ne disposions toujours pas de mesure de similarité de Relations distincte d'une mesure de $sim_a$. Nous avons cependant décidé de nous approprier ce que nous appellerons désormais *score de Turney*.

$$score_{Turney}(A : B :: C : D) = \frac{1}{2}(sim_a(A, C) + sim_a(B, D))$$

Nous faisions alors l'hypothèse qu'un tel score pouvait nous permettre d'évaluer la pertinence de différentes relations analogiques entre descriptions de lexies. Nous supposions également qu'il serait possible d'extraire des règles des analogies obtenant les scores les plus élevés.

### 3.1.4 Résumé des hypothèses

Les résultats fournis par notre prototype seront observés dans la section 3.5 au regard des 5 hypothèses résumées ici. Il s'agira d'en évaluer la pertinence.

HYPOTHÈSE 1 : Un degré de $sim_a$ élevé entre descriptions de lexies est un indice de Relations pertinentes entre leurs Attributs. Ces Relations peuvent être de trois ordres : Relations internes propres à chaque description, Relations externes communes aux deux descriptions dont le degré de $sim_a$ est élevé ou Relations externes communes à toutes les descriptions appartenant à un couple ayant un tel degré de $sim_a$.

HYPOTHÈSE 2 : La mesure de $sim_a$ entre descriptions de lexies appartenant à un même vocable fait émerger des Relations plus pertinentes qu'entre descriptions de lexies de vocables distincts. Il est ici question de Relations internes aux descriptions ou de Relations externes communes à l'intérieur du vocable.

HYPOTHÈSE 3 : Les descriptions de lexies d'un même vocable ayant un certain degré de $sim_a$ entrent en relation analogique avec les descriptions de lexies de tout autre vocable ayant entre elles une $sim_a$ analogue.

HYPOTHÈSE 4 : Les proportions analogiques permettant de faire émerger des règles pertinentes sont celles qui mettent en relation des DULB et des DULS de la manière suivante :
$DULS_{voc1} : DULB_{voc1} :: DULS_{voc2} : DULB_{voc2}$

HYPOTHÈSE 5 : Le calcul de score proposé par Turney (2006) peut être utilisé comme indice de qualité des proportions analogiques.

## 3.2 Sélection des données

Après avoir défini l'ensemble des hypothèses que nous souhaitions tester, notre attention s'est portée sur la sélection des données sur lesquelles les tester. Nous avons choisi de nous intéresser aux descriptions de lexies nominales du RL-fr ayant au minimum deux liens de FL sortants et appartenant à un vocable polysémique, c'est-à-dire contenant au moins deux lexies.

Par lexies nominales, nous entendons l'ensemble des lexies ayant N pour partie du discours profonde[10] (nom commun, nom propre, numéral, locution nominale, pronom, locution pronominale, pronom clitique, pronom personnel et pronom relatif).

Pour sélectionner ces descriptions, nous avons interrogé la base de données du RL-fr à l'aide d'une série de requêtes détaillées plus bas, en 3.4.

Après différentes observations partielles, les données utilisées correspondent aux descriptions obtenues le 30 juillet 2012. Il s'agit des descriptions de 118 lexies, réparties en 57 vocables.

Comme nous l'avons déjà évoqué, nous ne disposions alors, pour chacune de ces 118 lexies, que de trois types d'informations fiables, en ce sens qu'elles étaient d'ores et déjà entièrement gérées par l'éditeur Dicet à l'aide duquel les lexicographes tissent le réseau lexical et que leurs formes étaient normalisées[11]. Il s'agissait des CG, des FL et de l'appartenance à un vocable.

Nous allons à présent énumérer l'ensemble des CG et des FL utilisées pour la description de cette sélection de lexies. Nous détaillerons ensuite la réalisation technique de notre prototype.

### 3.2.1 Caractéristiques grammaticales

| CARACTÉRISTIQUES GRAMMATICALES FONDAMENTALES | |
|---|---|
| Partie du discours | |
| `nom commun` | 116 |
| `locution nominale` | 2 |
| Genre | |
| `fém` | 51 |
| `masc` | 64 |
| `masc, s'emploie aussi au fém` | 2 |
| FLEXION ET AUTRES CARACTÉRISTIQUES FORMELLES | |
| Détermination et nombre | |
| `pas de pl` | 3 |
| `pl` | 1 |
| `sg` | 1 |
| `surtout déf sg` %1 | 1 |
| MARQUES D'USAGE | |
| Marques d'usage langagier | |
| `rare` | 2 |

TAB. 3.2 : Caractéristiques grammaticales des données de l'expérience initiale

[10] Voir Mel'čuk (2006) pour une introduction à la notion de partie du discours profonde.

[11] Attention, si l'encodage de ces informations est techniquement fiable, nous ne présumons en rien de leur fiabilité lexicographique. Il est évidemment possible que certaines d'entre elles soient modifiées avant la publication des lexies.

Le tableau 3.2 énumère les CG présentes dans notre sous-ensemble de descriptions de lexies et précise le nombre d'occurrences de chacune. Nous pouvons y observer que les lexies collectées sont essentiellement des noms communs et que leurs descriptions mettent en jeu peu de caractéristiques en dehors des caractéristiques grammaticales fondamentales.

Notez également que nous y observons deux caractéristiques concurrentes, `pas de pl` et `sg`. Cette concurrence n'est pas motivée linguistiquement. Elle a, depuis, été éradiquée du RL-fr, où seule la caractéristique `sg` demeure.

### 3.2.2 Fonctions lexicales

Le tableau 3.3 fournit des informations sur les liens de FL présents dans les 118 descriptions de lexies dont nous disposions. Rappelons-le, les visées du regroupement en familles de FL qu'il propose sont avant tout pratiques pour leur énumération et non théoriques. Ce regroupement est emprunté à Mel'čuk et al. (1995) et à la base de FL encapsulée dans le RL-fr.

| **Familles de FL** | | **Liens** | **FL** |
|---|---|---|---|
| PARADIGMATIQUES | | 482 | 63 FL≠ |
| Synonymes | **Syn** | 41 | **Syn** |
| | | 40 | **Syn**$_{\cap}$ |
| | | 1 | **Syn**$_{\subset}$ |
| | | 6 | **Syn**$_{\supset}$ |
| | | 27 | **Syn**$_{\cap \mathbf{sexe}}$ |
| | | 4 | **Hypo** |
| Antonymes | **Anti** | 13 | **Anti** |
| | | 3 | **Anti**$_{\cap}$ |
| | | 5 | **Anti**$_{\supset}$ |
| | **AntiMagn** | 2 | **AntiMagn** |
| Contrastifs | **Contr** | 3 | **Contr** |
| Génériques | **Gener** | 8 | **Gener** |
| Dérivés syntaxiques | **S$_0$** | 2 | **S$_0$** |
| | | 1 | **S$_0$Pred** |
| ... | | | |

| | | | |
|---|---|---|---|
| ... | | | |
| | | 1 | $\mathbf{S_0Pred_{\subset}}$ |
| | $\mathbf{V_0}$ | 28 | $\mathbf{V_0}$ |
| | | 1 | $\mathbf{V_{0\cap}}$ |
| | | 2 | **Enun** |
| | $\mathbf{A_0}$ | 28 | $\mathbf{A_0}$ |
| | | 6 | $\mathbf{A_1 \vee A_0}$ |
| | $\mathbf{Adv_0}$ | 6 | $\mathbf{Adv_0}$ |
| Dérivés sémantiques nominaux actanciels | $\mathbf{S_1}$ | 62 | $\mathbf{S_1}$ |
| | | 4 | $\mathbf{S_{1/2}}$ |
| | | 2 | $\mathbf{S_{1/2}Pref}$ |
| | | 22 | $\mathbf{S_1}^{\mathbf{usual}}$ |
| | | 2 | **individu** $\mathbf{S_1}$ |
| | | 1 | $\mathbf{MultS_1}$ |
| | | 2 | **type particulier** $\mathbf{S_1}$ |
| | | 1 | $\mathbf{S_3 \vee S_1}$ |
| | $\mathbf{S_2}$ | 38 | $\mathbf{S_2}$ |
| | | 2 | $\mathbf{S_{1\supset}}$ |
| | | 2 | $\mathbf{S_{1\cap}}$ |
| | | 2 | $\mathbf{S_{loc}\&S_2}$ |
| | | 16 | **type particulier** $\mathbf{S_2}$ |
| | | 3 | **fait** $\mathbf{S_2}$ |
| | | 3 | **individu** $\mathbf{S_2}$ |
| | | 3 | $\mathbf{SingS_2}$ |
| | | 2 | $\mathbf{MultS_2}$ |
| | $\mathbf{S_3}$ | 21 | $\mathbf{S_3}$ |
| | $\mathbf{S_4}$ | 3 | $\mathbf{S_4}$ |
| | | 3 | $\mathbf{S_{loc}\&S_4}$ |
| Dérivés sémantiques nominaux circonstanciels | | 1 | $\mathbf{S_{instr}}$ |
| | | 1 | $\mathbf{S_{med}}$ |
| ... | | | |

| | | | |
|---|---|---|---|
| ... | | | |
| | | 1 | $\mathbf{S_{res}}$ |
| Singulatifs | | 4 | **Sing** |
| | | 3 | pour la vente **Sing** |
| | | 3 | fonctionnel, intérieur **Mero** |
| | | 1 | structural **Mero** |
| Collectifs | | 2 | **Mult** |
| | | 1 | **Mult**$\supset$ |
| | | 3 | pour la vente **Mult** |
| Nom de chef | | 1 | **Cap** |
| Nom d'équipe | | 1 | **Equip** |
| Dérivés sémantiques adjectivaux actanciels | $\mathbf{A_1}$ | 19 | $\mathbf{A_1}$ |
| | | 2 | $\mathbf{A_{1\supset}}$ |
| | | 1 | $\mathbf{Able_1 Caus_1 Manif_1}$ $\vee$ $\mathbf{A_1 Caus_1 Manif}$ |
| | | 2 | **Magn** + $\mathbf{A_1}$ |
| | | 5 | $\mathbf{A_{1/2}}$ |
| | | 1 | $\mathbf{S_0 A_1}$ |
| | | 1 | $\mathbf{A_{1/2}}$**Perf** |
| | $\mathbf{A_2}$ | 3 | $\mathbf{A_2}$ |
| Dérivés sémantiques adjectivaux potentiels | $\mathbf{Able_1}$ | 1 | $\mathbf{Able_1}$ |
| | $\mathbf{Able_2}$ | 4 | $\mathbf{Able_2}$ |
| SYNTAGMATIQUES | | 105 | 39 FL$\neq$ |
| FL adjectivales | | | |
| Intensificateurs | **Magn** | 14 | **Magn** |
| | | 1 | $\mathbf{Magn^{temp}}$ |
| | | 2 | $\mathbf{Magn_2}^{\mathbf{quant}}$ |
| | | 2 | $\mathbf{Magn}_{\mathbf{manifestation}}$ |
| | | 1 | **AntiBon** + **Magn** |
| | | | ... |

| | | | |
|---|---|---|---|
| ... | | | |
| « Confirmateurs » | **Ver** | 3 | **Ver** |
| FL adverbiales | | | |
| Dérivés sémantiques adverbiaux actanciels | **$Adv_1$** | 2 | **$Adv_1$** |
| | | 2 | **$Adv_1 Caus_1 Manif_1$** |
| Locatifs | **Loc** | 3 | **$Loc_{in}$** |
| FL verbales | | | |
| Verbes supports | **Oper** | 7 | **$Oper_1$** |
| | | 1 | **$Oper_{1+2}$** |
| | | 1 | **$Caus_2 Oper_1$** |
| | | 1 | **$Oper_2$** |
| | | 6 | **$Oper_{23}$** |
| | | 3 | **$Oper_3$** |
| | **Func** | 2 | **$Func_0$** |
| | | 1 | **$IncepFunc_0$** |
| | | 5 | **$CausFunc_0$** |
| | | 1 | **$Func_2$** |
| | | 1 | **$CausFunc_2$** |
| | | 1 | **$Caus_2 Func_2$** |
| | **Labor** | 1 | **$Labor_{21}$** |
| Verbes de réalisation | **Real** | 11 | **$Real_1$** |
| | | 1 | **$S_0 Real_1$** |
| | | 1 | **$S_0 SingReal_1$** |
| | | 1 | **$Real_2$** |
| | | 3 | **$Real_2^{\ I}$** |
| | | 2 | **$Real_2^{\ II}$** |
| | | 3 | **$Prepar_2 Real_2$** |
| | | 1 | **$AntiReal_2^{\ I}$** |
| | | 1 | **$AntiReal_2^{\ I/II}$** |
| | | 1 | **$AntiReal_2^{\ II}$** |
| ... | | | |

| | | | |
|---|---|---|---|
| ... | | | |
| | **Fact** | 1 | **$Fact_0$** |
| | | 3 | **déplacement $Fact_0$** |
| | | 1 | **$IncepFact_0$** |
| | | 3 | **$Caus_1Fact_0$** |
| | | 4 | **$Liqu_3Fact_0$** |
| | | 7 | **$Fact_2$** |
| Verbes causatifs | **Caus** | 1 | **$Caus_{1/2}$** |

TAB. 3.3 : Liens de FL des données de l'expérience initiale.

Lors de la sélection des données, certaines lexies du RL-fr avaient déjà été travaillées en profondeur et comportaient de nombreux liens de FL. D'autres, en revanche, ne contenaient que les liens de FL sortants encodés lors de la phase de constitution de la nomenclature directement induite. Ainsi la lexie ENNEMI$_N$ **I** comportait trente liens de FL sortants, tandis que près de 35% des lexies n'en comportaient que deux.

Le tableau 3.4 présente la répartition du nombre de liens de FL sortants par description de lexie.

| Nbr de liens de FL | Nbr de lexies concernées | % de lexies concernées |
|---|---|---|
| 2 | 41 | 34,75% |
| 3 | 25 | 21,19% |
| 4 | 10 | 8,47% |
| 5 | 16 | 13,56% |
| de 6 à 10 | 15 | 12,71% |
| de 11 à 20 | 6 | 5,08% |
| de 21 à 30 | 5 | 4,24% |

TAB. 3.4 : Répartition du nombre de liens de FL sortants par lexie

## 3.3 Format de représentation

Passons à présent aux aspects techniques de notre expérience. Comme nous l'avons indiqué en 3.1.1, nous avons effectué cette expérience en considérant les descriptions des lexies du RL-fr comme des objets proches de chaînes de caractères. Nous nous sommes donc basée sur l'analogie suivante :

la caractéristique est à la description d'une lexie
ce que
le caractère est à la chaîne de caractères

Le format de données utilisé devait donc nous permettre d'isoler deux types d'unités : les descriptions de lexie et les caractéristiques qui les composent. Une répartition des descriptions par fichier (1 fichier = 1 lexie) et des caractéristiques par ligne (1 ligne = 1 caractéristique) offre cette possibilité.

### 3.3.1 Notation RDF N3

L'une des particularités du RL-fr est d'être structuré en graphe. Même si nous avions décidé de mettre à plat sa structure dans le cadre de cette première expérience, nous avions eu par ailleurs l'occasion de nous intéresser aux standards de représentation de telles données. Parmi ceux-ci, le cas de RDF[12] a particulièrement retenu notre attention. RDF est un standard de représentation de données organisé en triplets [**sujet**, **prédicat**, **objet**] :

- le **sujet** représente une ressource à décrire ;
- le **prédicat** représente un type de propriété applicable au sujet ;
- l'**objet** représente une donnée ou une autre ressource, valeur de la propriété appliquée au sujet.

La notation 3, ou N3, est une norme de codage des modèles RDF sous forme textuelle dans laquelle les triplets sont séparés par des points. Ce format offre l'avantage d'être simple et rigoureux. L'organisation des informations qu'il propose nous a semblé répondre à nos besoins.

### 3.3.2 Éléments de description des lexies

Comme nous l'avons précédemment évoqué, nous disposions de trois types d'information fiable pour la réalisation de notre première expérience. Il s'agissait de l'appartenance d'une lexie à un vocable, de ses CG et des liens de FL qu'elle entretenait avec les autres lexies. Nous désignerons ici ces trois types d'information sous le terme d'*éléments de description*. Nous distinguerons ces éléments de description de leur réalisation sous forme de triplets [sujet, prédicat, objet] constituant les lignes de nos fichiers. De tels triplets seront désignés sous le terme de *caractéristiques*.

Chaque unité d'information contenue dans le RL-fr dispose à la fois d'un nom et d'un identifiant unique, comme le présente le tableau 3.5 pour les CG de genre.

| Nom | ID |
|---|---|
| `fém` | 26 |
| `fém, s'emploie aussi au masc` | 27 |
| `masc` | 28 |
| `masc, s'emploie aussi au fém` | 29 |

TAB. 3.5 : Désignation des CG de genre dans le RL-fr

Nous avons choisi d'utiliser les identifiants uniques dans la constitution des caractéristiques. Afin d'éviter toute ambiguïté, ils ont été préfixés en fonction de la nature

[12]Resource Description Framework `http://www.w3.org/RDF/`

des unités qu'ils identifiaient : V pour vocable, L pour lexie, CG pour caractéristique grammaticale, FL pour fonction lexicale .

### Appartenance de la lexie décrite à un vocable

L'appartenance de la lexie décrite à un vocable est un élément de description simple. Nous avons considéré que nous pouvions traduire cet élément en une seule caractéristique, par l'intermédiaire d'un prédicat « rlf :hasVoc » et de l'identifiant unique de chaque vocable :

```
rlf:lexie rlf:hasVoc rlf:V31447.
```

### Caractéristiques grammaticales

Les CG sont des éléments de descriptions simples. En langage courant, il est possible de les énumérer facilement à l'aide du verbe « être », comme dans la phrase « La lexie CORRESPONDANCE **II.a est** un nom commun. »[13].

Dans la norme de codage N3, il existe un mot-clef utilisable comme prédicat dans ce genre de situation. Il s'agit du mot-clef **a**, que nous avons décidé d'utiliser :

```
rlf:lexie a rlf:CG20.
rlf:lexie a rlf:CG26.
```

### Liens de fonctions lexicales

Pour représenter la description de la lexie CORRESPONDANCE **II.a** [*Les photos de famille circulent par tous les moyens de correspondance moderne en fichiers attachés*], il faut rendre compte du lien de FL qu'il existe entre cette lexie et la lexie LETTRE (L26941), par application de la FL de nominalisation du troisième actant sémantique $\mathbf{S_3}$ (FL59).

Cet élément complexe peut se représenter en considérant la lexie courante comme **sujet**, la FL comme **prédicat** et la valeur d'application de la FL comme **objet** :

```
rlf:lexie rlf:FL59 rlf:L26941.
```

Mais si nous avions opté pour une telle notation, aucune autre lexie n'aurait partagé cette caractéristique. Nous avons donc préféré répartir cet élément de description complexe en trois caractéristiques simples :

- une FL est en jeu dans la description courante : `rlf:FL59 a rlf:FL.`
- il existe un lien d'application de cette FL, ayant comme source la lexie courante : `rlf:lexie rlf:isArg rlf:FL59.`
- la cible de ce lien est elle-même une lexie : `rlf:FL59 rlf:value rlf:L26941.`

Dans cette représentation, le cas de FL dont la valeur d'application est un ensemble de plus d'une lexie peut être traité de la manière suivante : la première

[13]Dans le cas de certaines caractéristiques, la formulation en langage courant à l'aide du verbe « être » n'est pas idéale, comme pour la caractéristique `emploi adverbial`.

caractéristique reste unique, tandis que les deux autres sont multipliées.

Nous disposons alors de trois points de comparaison :

- mettre en œuvre la même FL dans sa description ;
- avoir le même nombre de liens de FL ;
- avoir des cibles communes par application d'une même FL.

La figure 3.1 montre la description de la lexie CORRESPONDANCE **II.a** dans un tel format. Les deux premières lignes, dont il n'a pas encore été question, sont communes à tous les fichiers et sont propres au format RDF N3. La troisième ligne, dont nous n'avons pas non plus parlé, contient le nom complet de la lexie courante. Elle est propre à chaque fichier. Elle est sans pertinence dans le cadre de la comparaison de descriptions de lexie, mais a facilité notre lecture lors de l'analyse des résultats.

```
@prefix rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix rlf:<http://relief.atilf.fr/RLF#>.
rlf:lexie rlf:name rlf:correspondance---II.a.
rlf:lexie rlf:hasVoc rlf:V31447.
rlf:lexie a rlf:CG20.
rlf:lexie a rlf:CG26.
rlf:FL23 a rlf:FL.
rlf:FL34 a rlf:FL.
rlf:FL59 a rlf:FL.
rlf:lexie rlf:isArg rlf:FL23.
rlf:lexie rlf:isArg rlf:FL34.
rlf:lexie rlf:isArg rlf:FL34.
rlf:lexie rlf:isArg rlf:FL59.
rlf:lexie rlf:isArg rlf:FL59.
rlf:FL23 rlf:value rlf:L32847.
rlf:FL34 rlf:value rlf:L32851.
rlf:FL34 rlf:value rlf:L32850.
rlf:FL59 rlf:value rlf:L31779.
rlf:FL59 rlf:value rlf:L26941.
```

FIG. 3.1 : Fichier RDF N3 de la lexie CORRESPONDANCE **II.a**

La figure 3.2 montre un extrait de la description lexicographique de cette lexie dans l'éditeur lexicographique Dicet.

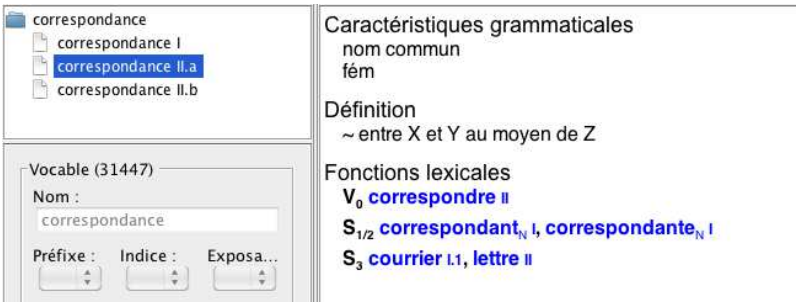



FIG. 3.2 : Vue de la lexie CORRESPONDANCE **II.a** dans l'éditeur Dicet

## 3.4 Implémentation du prototype

Nous avons jusqu'à présent détaillé les hypothèses que nous souhaitions tester, les données choisies à cette fin et leur format de représentation. Nous allons maintenant faire le point sur la mise en place technique de notre prototype.

### 3.4.1 Extraction des données

Avant toute chose, nous avons effectué un export de la base de données du RL-fr afin d'interroger localement les données et d'effectuer des manipulations. Nous avons été amenée à créer des tables virtuelles issues de l'assemblage de tables SQL existantes et ne souhaitions pas modifier la base de travail des lexicographes. De plus, le RL-fr évolue continuellement et nous souhaitions pouvoir consulter à tout moment l'état correspondant aux données de notre expérience.

#### 3.4.1.1 Identification des lexies

Une fois la base importée sur notre serveur local, nous avons effectué un ensemble de requêtes SQL pour obtenir la liste des lexies répondant à nos critères. Comme nous l'avons dit précédemment, il s'agissait des lexies nominales, ayant un minimum de 2 liens de FL sortants et appartenant à des vocables polysémiques.

Le tableau 3.6 présente le nombre de résultats retournés pour chacune des requêtes effectuées. Nous avons finalement abouti à une liste de 118 identifiants de lexies répondant à nos critères de sélection.

| Objet de la requête | Nbr de résultats |
|---|---|
| Vocables comportant un minimum de 2 lexies | 1406 |
| Lexies ayant un minimum de 2 liens de FL sortant | 1827 |
| Lexies nominales | 6015 |
| Lexies nominales ayant un minimum de 2 liens de FL sortant | 954 |
| Vocables des lexies nominales ayant un minimum de 2 liens de FL sortant | 57 |
| Lexies nominales, ayant un minimum de 2 liens de FL sortant et appartenant à des vocables comportant au moins 2 lexies | 118 |

TAB. 3.6 : Requêtes ayant permis la sélection des données

#### 3.4.1.2 Extraction des descriptions

Dans un second temps, nous avons développé un script PHP permettant d'obtenir un fichier de description pour chacune des 118 lexies sélectionnées.

Ce script interroge notre copie locale de la base de données du RL-fr et collecte les informations contenues dans chaque description de lexie. Un fichier d'extension

`n3`, conforme au format décrit dans la section 3.3, a alors été généré par description.

Les noms des fichiers se composent de l'identifiant unique du vocable suivi de l'identifiant unique de la lexie, selon le modèle `idVocable_idLexie.n3`.

### 3.4.2 Mesures de similarité

Une fois les différents fichiers de descriptions de lexies obtenus, nous avons mis au point un script PERL permettant le calcul d'une distance d'édition, puis un script BASH permettant d'appliquer ce calcul aux descriptions de lexies.

Le résultat de ces opérations se présente sous la forme d'un fichier `csv` à cinq colonnes dont la figure 3.3 présente un aperçu. Les descriptions de lexies y sont identifiées selon le même modèle que les noms des fichiers RDF N3, `idVocable_idLexie`.

| lexie1 | lexie2 | distance d'édition | dissemblance | similarité |
|---|---|---|---|---|
| 26230_26228 | 26230_26228 | 0 | 0 | 1 |
| 26230_26228 | 26230_34245 | 6 | 0.0833333333333333 | 0.916666666666667 |
| 26230_26228 | 26242_26240 | 63 | 0.777777777777778 | 0.222222222222222 |
| 26230_26228 | 26242_29964 | 54 | 0.75 | 0.25 |
| 26230_26228 | 26405_26403 | 198 | 0.916666666666667 | 0.0833333333333334 |
| 26230_26228 | 26405_34059 | 63 | 0.777777777777778 | 0.222222222222222 |

FIG. 3.3 : Exemple de sortie du script de calcul de distance d'édition

#### 3.4.2.1 Calcul de distance d'édition

Notre script PERL de calcul d'édition est une adaptation de l'algorithme de Wagner et Fischer (1974). Il prend en entrée deux fichiers `idVocable_idLexie.n3`. Le calcul de distance d'édition qu'il utilise est conforme au calcul de distance d'édition canonique entre chaînes de caractères présenté en 3.1.1.1 et obtenu à l'aide d'une matrice.

Il calcule également les mesures de dissemblance et de $sim_a$ d'après les propositions de Lepage (2003) présentées en 3.1.1.2.

En parallèle du fichier `csv`, un fichier trace est généré. Il contient le détail des comparaisons ainsi que la matrice complète de calcul d'édition canonique.

Pour simplifier notre propos, supposons que nous nous intéressions à la distance entre les chaînes de caractères **tua** et **use**. Notre script construirait alors une matrice de cinq lignes sur cinq colonnes, telle que celle présentée dans le tableau 3.7[14].

La première ligne de cette matrice est consacrée à la chaîne de caractères **tua**. Elle commence par une case vide, suivie d'une case contenant le caractère vide (que nous choisissons de symboliser à l'aide du caractère ⊘) et de trois cases contenant, l'un après l'autre, les caractères **t**, **u** et **a**.

[14]La première ligne et la première colonne sont données ici dans un but explicatif, mais ne sont pas nécessaires. Pour un tel cas, le script développé construirait en réalité une matrice de taille 4x4.

| | ⊘ | **t** | **u** | **a** |
|---|---|---|---|---|
| ⊘ | 0 | 1 | 2 | 3 |
| **u** | 1 | 2 | 1 | 3 |
| **s** | 2 | 3 | 2 | 3 |
| **e** | 3 | 4 | 3 | **4** |

TAB. 3.7 : Matrice de calcul de distance d'édition canonique entre **tua** et **usa**

La première colonne est, elle, consacrée à la chaîne de caractères **use**. Elle commence par une case vide, suivie d'une case contenant le caractère vide et de trois cases contenant, l'un après l'autre, les caractères **u**, **s** et **e**.

La seconde ligne et la seconde colonne ne nécessitent aucun calcul, elles sont remplies en partant de 0 et en allant jusqu'à la longueur de la chaîne qui a déterminé leur longueur (**tua** pour la ligne, **use** pour la colonne).

Les autres cases sont remplies de la manière suivante :

- Le script compare les deux caractères qui se croisent en ce point de la matrice. S'ils sont identiques, la substitution de l'un par l'autre coûte 0, s'ils sont différents, elle coûte 2. Ce coût est ajouté à la valeur de la case située en diagonale en haut à gauche de la case courante pour obtenir la valeur de la case courante en cas de substitution.
- Le coût de l'effacement est de 1. Ce coût est ajouté à la valeur de la case située à gauche de la case courante pour obtenir la valeur de la case courante en cas d'effacement.
- Le coût de l'insertion est de 1. Ce coût est ajouté à la valeur de la case située au-dessus de la case courante pour obtenir la valeur de la case courante en cas d'insertion.
- Les valeurs obtenues par les opérations de substitution, d'effacement et d'insertion sont comparées et la plus petite des trois est finalement attribuée à la case courante.

La distance d'édition canonique entre les deux chaînes se lit dans la dernière case de la dernière ligne de la matrice, ici 4.

La dissemblance est obtenue en divisant ce résultat par la longueur cumulée des chaînes **tua** et **use** : $\frac{4}{3+3} \simeq 0,67$.

La $sim_a$ est obtenue en soustrayant la dissemblance à 1 : $1 - 0,67 = 0,33$.

#### 3.4.2.2 Comparaison des descriptions de lexies

Afin de tester l'ensemble de nos hypothèses, nous avons effectué différentes séries de comparaisons de descriptions de lexies à l'aide de notre script BASH.

Une première série de calculs a consisté à comparer chaque description de lexie à l'ensemble des descriptions de lexies partageant le même vocable qu'elle. Le fichier `distance_edition_in_voc_3007.ods` répertorie les résultats de ces calculs.

Une seconde série de calculs a consisté à comparer chaque description de lexie à l'ensemble des autres descriptions de lexies. Elle aboutit à la création d'un second fichier, `distance_edition_all_3007.ods`[15]. Son objectif est notamment de vérifier la pertinence de notre deuxième hypothèse[16].

Ces deux séries de calculs nous ont permis de constater qu'une faible proportion de mesure de $sim_a$ aboutit à un résultat élevé, comme le montre le tableau 3.8.

| | total $sim_a$ | nbr $\geq 0,75$ | % $\geq 0,75$ | nbr $\geq 0,5$ | % $\geq 0,5$ |
|---|---|---|---|---|---|
| all | 6903 | 8 | 0,116 | 159 | 2,303 |
| in voc | 68 | 6 | 8,823 | 23 | 33,823 |

TAB. 3.8 : Répartition des similarités d'Attributs

Différentes autres séries de calculs ont ensuite été effectuées pour observer des comparaisons plus restreintes, comme les couples de descriptions de lexies d'un même vocable ayant une $sim_a$ supérieure à 0,75 ou située entre 0,50 et 0,75, les DULB de ces couples entre elles, ou encore les DULS de ces couples.

### 3.4.3 Score de Turney

En dernier lieu, nous avons ajouté à notre script BASH une fonctionnalité permettant le calcul de scores de Turney de proportions analogiques construites à partir des couples de $sim_a \geq 0,75$ d'une part et comprise entre $0,5$ et $0,75$ d'autre part.

Dans le cas de la comparaison de chaque description de lexie à l'ensemble des descriptions de lexies partageant le même vocable, les proportions analogiques construites respectent le modèle $DULS_{voc1} : DULB_{voc1} :: DULS_{voc2} : DULB_{voc2}$.

Chaque couple ($DULS_{voc1}$, $DULS_{voc2}$) et ($DULB_{voc1}$, $DULB_{voc2}$) en œuvre dans ses proportions est envoyé au script de calcul de distance d'édition. Le score de Turney est alors le résultat de l'addition de ces deux $sim_a$, divisé par 2. Le tableau 3.9 présente une répartition des résultats de ce calcul.

Nous constatons d'ores et déjà que peu de proportions ainsi constituées obtiennent un score élevé.

[15]Ce fichier contient notamment la $sim_a$ entre une lexie et elle-même et la $sim_a$ entre la description de la lexie A et la description de la lexie B, ainsi qu'entre la description de la lexie B et la description de la lexie A, ce qui n'est pas pertinent. Les données numériques présentées ici tiennent compte de cela et sont à comprendre sans les 118 comparaisons d'une description avec elle-même et sans doublon.

[16]Il est évident que l'ensemble des mesures obtenues lors des autres séries de calculs est également disponible dans le fichier `distance_edition_all_3007.ods`, mais sa taille en rend la lecture et la manipulation peu aisées.

| | total *scores* | $scores \geq 0,75$ | $scores \geq 0,5$ | $scores \geq 0,4$ |
|---|---|---|---|---|
| $sim_a \geq 0,75$ | 15 | 0 | 2 | 4 |
| $0,5 \leq sim_a < 0,75$ | 136 | 0 | 5 | 21 |

TAB. 3.9 : Répartition des scores de Turney des proportions analogiques de type $DULS_{voc1} : DULB_{voc1} :: DULS_{voc2} : DULB_{voc2}$

### 3.4.4 Résumé des fichiers disponibles

Pour analyser les données et tester nos hypothèses, nous disposions donc des six fichiers suivants :

1. un fichier `idVocable_idLexie.n3` par description de lexie ;
2. un fichier `distance_edition_all_3007.ods` qui contenait l'ensemble des calculs de $sim_a$ possibles entre descriptions de lexies ;
3. un fichier `distance_edition_in_voc_3007.ods`, qui contenait les calculs de $sim_a$ entre descriptions de lexies appartenant à un même vocable ;
4. un fichier `lexie1_vs_lexie2` par calcul de distance d'édition canonique, qui contenait la trace de ce calcul ;
5. un fichier `calcul_scores.ods` pour les proportions analogiques construites avec les couples de descriptions de lexies de même vocable de $sim_a \geq 0,75$, qui contenait la mesure de score de Turney ;
6. un fichier `calcul_scores.ods` pour les proportions analogiques construites avec les couples de descriptions de lexies de même vocable de $sim_a$ comprise entre $0,5$ et $0,74$, qui contenait la mesure de score de Turney.

## 3.5 Analyse des résultats

La quantité de données disponible en l'état du RL-fr au 30 juillet 2012, si elle nous a semblé suffisante pour tester les différentes hypothèses présentées en 3.1, ne l'était sans doute pas pour permettre l'émergence de règles par analogie. Cette partie sera donc principalement consacrée à la confrontation des hypothèses aux données collectées, l'extraction de règles pertinentes et les moyens concrets de sa réalisation restant pour l'instant inexplorés.

### 3.5.1 Rappel concernant les données

L'analyse des résultats obtenus doit se faire en gardant à l'esprit quelques particularités des données disponibles.

**A)** Il s'agit de descriptions de lexies nominales, appartenant à des vocables polysémiques ayant au moins deux liens de FL sortants.

**B)** Ces descriptions sont réparties en caractéristiques, que nous considérons comme des Attributs.

Nous distinguons principalement deux types de caractéristiques : celles relatives aux CG et celles relatives aux liens de FL sortants. Les liens de FL entrants ne sont pas pris en compte.

**C)** Aucun poids n'est attribué aux différentes caractéristiques.

Cependant, nous garderons à l'esprit que la correspondance de partie du discours et de genre est une configuration normale dans le cas de descriptions de lexies appartenant à un même vocable.

**D)** Les liens de FL sont des caractéristiques complexes, mettant en œuvre une FL, l'application de cette FL à la lexie décrite et la valeur d'application de cette FL.

Dans le format de représentation de données choisi, un lien de FL correspond à trois caractéristiques distinctes : une occurrence de FL en jeu dans la description, une application de la FL à la lexie en cours de description et une valeur d'application, aussi appelée cible du lien de FL. Si l'application d'une FL à une lexie aboutit à une valeur comportant un ensemble de plus d'un élément, nous disposons d'autant de triplets [FL en jeu, application de FL, cible du lien de FL] que d'éléments constituant la valeur.

**E)** En l'état du RL-fr au 30 juillet 2012, certaines descriptions de lexies étaient plus fournies en liens de FL que d'autres, comme le montre le tableau 3.4 de la section 3.2.

Une grande partie du travail lexicographique de tissage de liens de FL effectué jusqu'à cette date l'a été sur les DULB des vocables. Un déséquilibre en nombre de liens de FL et de caractéristiques correspondantes entre les DULB et les DULS est donc naturel.

Nous allons à présent confronter chacune de nos cinq hypothèses, présentées en 3.1, aux résultats fournis par notre prototype.

### 3.5.2 La similarité d'Attributs comme indice de similarité de Relations

> HYPOTHÈSE 1 : Un degré de $sim_a$ élevé entre descriptions de lexies est un indice de Relations pertinentes entre leurs Attributs. Ces Relations peuvent être de trois ordres : Relations internes propres à chaque description, Relations externes communes aux deux descriptions dont le degré de $sim_a$ est élevé ou Relations externes communes à toutes les descriptions appartenant à un couple ayant un tel degré de $sim_a$.

#### 3.5.2.1 Proportion de $sim_a$ élevées

La première chose que nous avons observée, c'est qu'une très faible proportion (2,07%) de calculs de $sim_a$ — pour l'ensemble des couples de descriptions de lexies possibles — aboutissait à une valeur supérieure ou égale à 0,5 et que celle des calculs

| Lexie $desc_1$ | Lexie $desc_2$ | $sim_a(desc_1, desc_2)$ |
|---|---|---|
| ÂGE **I.a** | ÂGE **I.b** | 0,917 |
| BOUCHER$_N$ **2** | BOULANGER$_N$ **2** | 0,75 |
| ACCOMPAGNATEUR **I** | ACCOMPAGNATEUR **II** | 0,75 |
| ACCOMPAGNATRICE **I** | ACCOMPAGNATRICE **II** | 0,75 |
| ACCENTUATION **I.1** | ACCENTUATION **I.2** | 0,75 |
| INCONSCIENCE **I** | INCONSCIENCE **II** | 0,75 |
| INCONSCIENCE **II** | CONSCIENCE **II** | 0,75 |
| CONSCIENCE **I** | CONSCIENCE **II** | 0,75 |

TAB. 3.10 : Similarité d'Attributs $\geq 0,75$

aboutissant à une valeur supérieure ou égale à 0,75 était portion congrue (0,116%).

En nous intéressant aux huit cas de $sim_a \geq 0,75$ présentés dans le tableau 3.10, nous avons constaté que seulement deux des couples de descriptions comparées obtenant un tel degré de $sim_a$ étaient composées de descriptions de lexies appartenant à des vocables distincts[17].

Le premier de ces couples de descriptions concerne les lexies BOUCHER$_N$ **2** [*Je vais chez le boucher acheter de la viande.*] et BOULANGER$_N$ **2** [*Chaque quartier a son boucher, son boulanger, son épicier*]. Il s'agit de DULS de vocables dont la lexie de base désigne un nom de métier. Cette caractéristique commune nous a semblé pertinente pour la détection de descriptions de lexies analogues.

Le second de ces couples concerne les lexies INCONSCIENCE **II** [*Comme trop souvent, elle avait agi avec inconscience, frivolité, égoïsme.*] et CONSCIENCE **II** [*Il est tant que tu prennes conscience de ton âge*]. Il s'agit de lexies antonymes. Cette particularité nous a amené à penser que leur rapprochement était également pertinent.

Un troisième cas a attiré notre attention, celui des descriptions de lexies du vocable ÂGE, dont le degré de $sim_a$ était de 0,917. En regardant en détail ces descriptions, nous avons constaté qu'elles étaient identiques en termes de CG, de nombre et de type de FL en jeu dans leur description, ainsi que de cibles de liens de FL. Leur degré de $sim_a$ vient donc confirmer ce que la numérotation de ces lexies, **I.a** [*Elle a eu son premier scooter à l'âge de 17 ans.*] et **I.b** [*Il existe toute sorte de calcul pour estimer l'âge des animaux*], nous dit déjà sur leur proximité.

L'information fournie par la numérotation n'a pas du tout été prise en compte dans ce prototype. Les quelques essais que nous avons effectués pour les intégrer à la procédure n'ont pas été concluants. Le RL-fr s'est, depuis, muni de liens de copolysémie, qui seront pris en compte dans la suite de nos travaux.

#### 3.5.2.2 Relations internes propres à chaque lexie

Pour savoir s'il existait des Relations pertinentes à l'intérieur des descriptions de lexies, nous avons choisi de nous concentrer sur leurs Attributs **FL en jeu dans**

[17] Nous nous garderons toutefois de tirer des conclusions sur une si petite quantité de données.

**la description courante**.

Sur les quatorze descriptions appartenant à des couples de $sim_a \geq 0,75$, seulement quatre nous ont permis de dégager des Relations intéressantes. Le tableau 3.11 présente les FL en jeu dans ces descriptions. Nous y constatons que les descriptions des lexies ÂGE **I.a** et ÂGE **I.b** ne mettent en jeu que des FL de la famille des dérivations sémantiques adjectivales actancielles du 1er actant. Les lexies BOUCHER$_N$ **2** et BOULANGER$_N$ **2**, pour leur part, ne mettent en jeu que des FL dont les valeurs d'application sont nominales.

| Lexie | FL en jeu dans la description |
|---|---|
| ÂGE **I.a** | $\mathbf{A_1}$<br>**Magn** + $\mathbf{A_1}$ |
| ÂGE **I.b** | $\mathbf{A_1}$<br>**Magn** + $\mathbf{A_1}$ |
| BOUCHER$_N$ **2** | $\mathbf{Syn}_{\cap\ \mathbf{sexe}}$<br>$\mathbf{type\ particulier}\ \mathbf{S_2}$<br>$\mathbf{S_3}$<br>$\mathbf{S_{loc}}\&\mathbf{S_4}$ |
| BOULANGER$_N$ **2** | $\mathbf{Syn}_{\cap\ \mathbf{sexe}}$<br>$\mathbf{type\ particulier}\ \mathbf{S_2}$<br>$\mathbf{S_3}$<br>$\mathbf{S_{loc}}\&\mathbf{S_4}$ |

TAB. 3.11 : Relations pertinentes entre FL en jeu dans la description des lexies appartenant à des couples de $sim_a \geq 0,75$

#### 3.5.2.3 Relations externes communes aux deux descriptions de lexies dont le degré de $sim_a$ est élevé

En nous intéressant aux Relations externes communes aux descriptions de chaque couple de $sim_a$ élevé, nous avons pu observer davantage de Relations qui nous ont semblé pertinentes.

Ainsi, sur les huit couples, toutes les descriptions comparées deux à deux partageaient les mêmes caractéristiques CG (en nombre et en valeurs) et des caractéristiques de liens de FL sortants identiques (en nombre et en terme de FL mises en jeu).

Les Relations communes observées étaient donc de type : **les CG sont au nombre de 2**, **les CG sont `nom commun` et `fém.`**, **les liens d'application de FL sont au nombre de 2**, **les FL en jeu sont $\mathbf{V_0}$ et $\mathbf{S_3}$**[18].

[18]Ces occurrences de Relations sont valables pour ACCENTUATION **I.1** et ACCENTUATION **I.2**.

#### 3.5.2.4 Relations externes communes à toutes les descriptions appartenant à un couple ayant un degré de $sim_a$ élevé

Enfin, en observant ensemble les quatorze descriptions, des associations privilégiées d'Attributs **FL en jeu dans leur description nous sont apparues**. Ces observations sont cependant à prendre avec précautions, vu le nombre réduit de lexies.

Le tableau 3.12 présente ces associations. Nous constatons ainsi que la présence de la FL **Fact$_2$** dans la description d'une lexie est systématiquement accompagnée de la présence de la FL **Syn$_{\cap\ \text{sexe}}$**[19], ce qui peut être analysé en lien avec la féminisation des noms de métiers. Le même phénomène est observé pour la FL **Anti**, dont la présence cooccurre toujours avec celle de la FL **A$_1$**.

| Occurrences de FL | | Cooccurrences de FL |
|---|---|---|
| **Syn$_{\cap\ \text{sexe}}$** | **Fact$_2$** | **Syn$_{\cap\ \text{sexe}}$** et **Fact$_2$** |
| 6 | 4 | 4 |
| **Anti** | **A$_1$** | **Anti** et **A$_1$** |
| 4 | 6 | 4 |

TAB. 3.12 : Cooccurrences récurrentes de FL en jeu dans la description des lexies appartenant à des couples de $sim_a \geq 0,75$

#### 3.5.2.5 Relations dans des couples de descriptions à faible $sim_a$

Pour compléter ces observations, nous avons regardé en détail deux couples de descriptions à faible $sim_a$. Il s'agissait du couple de descriptions des lexies EXCEPTION **1** [*Je dis bien en principe, car les exceptions sont très fréquentes.*] et CONSCIENCE **II** [*Il est tant que tu prennes conscience de ton âge*] d'une part, et des lexies ENNEMI$_N$ **I** [*Leurs relations avec le voisinage sont exécrables. Des ennemis partout.*] et ENNEMI$_N$ **III.b** [*Sous le feu des ennemis, j'ai traversé la Moselle à la nage.*] d'autre part.

$$sim_a(\text{EXCEPTION 1},\text{CONSCIENCE II}) \simeq 0,296$$

La description de la lexie EXCEPTION **1** contenait une Relation interne qui semblait intéressante. Les trois FL en jeu dans sa description (**V$_0$**, **A$_0$** et **Adv$_0$**), avaient toutes pour cibles des dérivés syntaxiques. De plus, ces dérivés couvraient les trois parties du discours adjectif, verbe et adverbe.

Cette Relation n'était pas partagée par CONSCIENCE **II**. Les seules Relations communes observées entre les deux lexies étaient **les CG sont au nombre de 2** et **les CG sont `nom commun` et `fém`**.

[19] La FL **Syn$_{\cap\ \text{sexe}}$** a fait l'objet d'un travail approfondi. Dans la version actuelle du RL-fr, cette FL a été remplacée par les FL **Syn$_{\supset}^{\text{sex}}$** et **Syn$_{\subset}^{\text{sex}}$**, que nous avons évoqués dans le chapitre 2. Delaite et Polguère (2013) discutent les raisons de cette modification.

$sim_a$(ENNEMI$_{\text{N}}$ **I**,ENNEMI$_{\text{N}}$ **III.b**) $\simeq 0,123$

La description de la lexie ENNEMI$_{\text{N}}$ **I** contenait une Relation interne qui pourrait éventuellement s'avérer intéressante. Ses Attributs **liens d'application de FL ayant pour source la lexie courante** se répartissaient en 9 liens mettant en jeu des FL ayant des cibles nominales — tissés à l'aide des FL **Syn$_{\cap\,\textbf{sexe}}$**, **Syn$_{\cap}$** et **Anti** — 10 liens mettant en jeu des FL ayant des cibles verbales — tissés à l'aide des FL **S$_0$Pred**, **S$_0$Pred$_{\subset}$**, **Oper$_{1+2}$**, **Oper$_1$**, **Oper$_2$**, **Caus$_2$Func$_2$**, **CausFunc$_2$**, **Labor$_{21}$** et **Fact$_2$** — et 11 liens mettant en jeu des FL ayant des cibles adjectivales — tissés à l'aide des FL **A$_0$**, **A$_1$**, **Magn$_{\textbf{manifestation}}$**, **Magn$^{\textbf{temp}}$** et **Magn**. Cette répartition est proche du 1/3, 1/3, 1/3 observé dans la description de EXCEPTION **1**.

Cette Relation n'était cependant pas partagée par la description de la lexie ENNEMI$_{\text{N}}$ **III.b**. Les seules Relations externes communes observées entre les deux descriptions étaient **les CG sont au nombre de 2** et **les CG sont `nom commun` et `masc`**. Ces Relations communes au sujet des CG sont d'autant moins pertinentes que les deux lexies partagent le même Attribut vocable.

#### 3.5.2.6 Conclusion

Le fait qu'une description appartienne à un couple ayant une $sim_a$ élevée ne semble pas être un indice de Relations pertinentes entre ses Attributs.

En revanche, cette description partage un certain nombre de Relations avec l'autre description du couple, qui peuvent s'avérer intéressantes. L'objectif de ce prototype étant la circonscription de relations analogiques et non synonymiques, cette constatation est plutôt encourageante.

L'observation de l'ensemble des lexies appartenant à des couples ayant une $sim_a$ élevée semble permettre de faire émerger des Relations intéressantes. Mais les données sont trop peu nombreuses pour valider cette hypothèse. De plus, seule une observation détaillée de l'ensemble des couples de descriptions, qu'ils appartiennent à des couples ayant une $sim_a$ faible ou élevée, permettrait une telle validation.

### 3.5.3 Le vocable comme espace privilégié

> HYPOTHÈSE 2 : La mesure de $sim_a$ entre descriptions de lexies appartenant à un même vocable fait émerger des Relations plus pertinentes qu'entre descriptions de lexies de vocables distincts.

#### 3.5.3.1 Proportion de $sim_a$ élevées entre lexies de même vocable

Nous avons observé précédemment que dans les huit couples de descriptions de $sim_a \geq 0,75$, seulement deux concernaient des lexies de vocables différents.

En nous intéressant aux mesures de $sim_a$ entre descriptions de lexies de même vocable, nous avons constaté que la proportion de $sim_a \geq 0,5$ était beaucoup plus élevée que pour l'ensemble des descriptions (33,823% contre 2,303%) et que celle de $sim_a \geq 0,75$ n'était plus aussi insignifiante que pour ce même ensemble (8,823%

contre 0,116%).

Ces observations sont conformes à ce que nous savons du regroupement de lexies au sein d'un même vocable. Il est en effet logique que les Attributs communs entre descriptions de lexies d'un même vocable soient plus fréquents qu'entre descriptions de lexies de vocables distincts. Notre mesure semble donc en adéquation avec ce que nous pouvons logiquement anticiper.

#### 3.5.3.2 Couples de descriptions hors vocable de $sim_a \geq 0,75$

En nous intéressant aux huit couples de descriptions de $sim_a \geq 0,75$, nous avons également observé que les descriptions des deux couples hors vocable étaient les seules, avec les descriptions des lexies du vocable ÂGE, à avoir des cibles de liens de FL communes, ce qui semblait être une ressemblance pertinente, bien qu'elle se rapproche de ce que nous attendons de lexies synonymes. La valeur d'application de FL commune dans les descriptions des lexies INCONSCIENCE **II** et CONSCIENCE **II** était en réalité une erreur, corrigée depuis. Le cas des lexies BOUCHER$_{\text{N}}$ **2** et BOULANGER$_{\text{N}}$ **2**, qui partageaient les cibles CLIENT et CLIENTE, était donc la seule pertinente.

De plus, les descriptions des lexies BOUCHER$_{\text{N}}$ **2** et BOULANGER$_{\text{N}}$ **2** étaient seules à partager la particularité interne de ne compter, dans leur description, que des cibles de FL nominales. À ce stade de l'analyse, l'hypothèse semblait donc erronée.

#### 3.5.3.3 Couples de descriptions hors vocables de $sim_a$ comprise entre 0,5 et 0,75

Pour confronter davantage notre hypothèse aux données, nous avons choisi de regarder en détail les couples de descriptions des lexies ACCOMPAGNEMENT **I** et ACCOMPAGNEMENT **III** de $sim_a \simeq 0,71$ d'une part, et des lexies AUTO **I** et VOITURE **1** de $sim_a \simeq 0,701$ d'autre part.

Le tableau 3.13 présente la comparaison des descriptions des lexies ACCOMPAGNEMENT **I** et ACCOMPAGNEMENT **III**. Nous pouvons y voir que les Relations qui se dégagent sont **2 des CG sont `nom commun` et `masc.`**, **les FL en jeu sont $\mathbf{V_0}$ et $\mathbf{S_1}$** et **les liens d'application de FL sont 1 lien mettant en jeu $\mathbf{V_0}$ et 2 mettant en jeu $\mathbf{S_0}$**. Par ailleurs, leur Attribut vocable est identique.

Les descriptions des lexies AUTO **I** et VOITURE **1** nous ont intéressée parce qu'il s'agit de lexies synonymes. Cette synonymie est actée dans leur description par la présence des liens de FL **Syn**(AUTO **I**) = (VOITURE **1**) et **Syn**(VOITURE **1**) = (AUTO **I**).

Le tableau 3.14 présente la comparaison des descriptions de ces deux lexies. Nous pouvons y voir se dégager les Relations **les CG sont au nombre de 2**, **les CG sont `nom commun` et `fém.`**, **les FL Syn, $\mathbf{A_0}$, $\mathbf{S_1}$, $\mathbf{S_{1/2}}$, $\mathbf{S_2}$, $\mathbf{Real_1}$ sont mises en jeu dans la description** et **le nombre de liens d'application de ces FL est identique**. Peut-être plus intéressant encore, treize de leurs Attributs **cibles de**

| | ACCOMPAGNEMENT I | ACCOMPAGNEMENT III |
|---|---|---|
| Nbr de CG | 3 | 2 |
| CG communes | `nom commun`, `masc.` | |
| Nbr de FL en jeu | 2 | 2 |
| FL communes | $\mathbf{V_0}$, $\mathbf{S_1}$ | |
| Nbr de liens de FL | 3 | 3 |
| Liens de FL communs | 1 $\mathbf{V_0}$, 2 $\mathbf{S_1}$ | |
| Nbr de cibles communes | 0 | |

TAB. 3.13 : Comparaison de ACCOMPAGNEMENT I et ACCOMPAGNEMENT III

**liens d'application de FL** sont identiques[20].

| | AUTO I | VOITURE 1 |
|---|---|---|
| Nbr de CG | 2 | 2 |
| CG communes | `nom commun`, `fém.` | |
| Nbr de FL en jeu | 10 | 7 |
| FL communes | **Syn**, $\mathbf{A_0}$, $\mathbf{S_1}$, $\mathbf{S_{1/2}}$, $\mathbf{S_2}$, $\mathbf{Real_1}$ | |
| Nbr de liens de FL | 18 | 15 |
| Liens de FL communs | 4 **Syn**, 2 $\mathbf{A_0}$, 4 $\mathbf{S_1}$, 1 $\mathbf{S_{1/2}}$, 2 $\mathbf{S_2}$, 1 $\mathbf{Real_1}$ | |
| Nbr de cibles communes | 13 | |

TAB. 3.14 : Comparaison de AUTO I et VOITURE 1

Ces observations nous ont confortée dans l'idée que les descriptions de lexies synonymes partagent un nombre important d'Attributs **cibles de liens d'application de FL**, en plus de Relations communes, et que la prise en compte de paires de lexies hors vocable nécessiterait un moyen de discriminer les synonymes.

#### 3.5.3.4 Conclusion

Il aurait été enrichissant de regarder en détail les cas de $sim_a$ comprises entre 0,5 et 0,75 pour approfondir l'analyse, mais ces observations ont déjà confirmé que l'hypothèse n'était pas exacte et qu'il serait intéressant d'inclure, dans la suite de la procédure de détection de relations analogiques, les couples de descriptions de lexies hors vocable partageant une $sim_a$ élevée.

[20]En réalité, quatorze d'entre elles sont identiques, mais la quatorzième est la valeur de $\mathbf{IncepFact_0}$(AUTO I) d'une part et de $\mathbf{Fact_0}$(VOITURE 1) d'autre part. Les lexies cibles étant associées aux liens de FL dans la représentation, cette dernière n'a pas pu être prise en compte pour le calcul de distance d'édition.

Cette inclusion ne pourra cependant pas se faire sans réfléchir davantage à la notion de synonymie et nécessitera sans doute la prise en compte de la notion de contiguïté.

### 3.5.4 Une similarité d'Attributs égale induit l'analogie

> HYPOTHÈSE 3 : Les descriptions de lexies d'un même vocable ayant un certain degré de $sim_a$ entrent en relation analogique avec les descriptions de lexies de tout autre vocable ayant entre elles une $sim_a$ analogue.

#### 3.5.4.1 Construction de proportions analogiques

Pour vérifier cette hypothèse, nous avons choisi de construire manuellement deux proportions analogiques à partir de couples de $sim_a = 0,75$ mis en évidence en 3.5.2 et une troisième à partir de couples de $sim_a \leq 0,20$.

Selon nos hypothèses, la validité des proportions analogiques dépend de l'existence de similarités de Relations entre les couples de descriptions qu'elles mettent en relation.

#### 3.5.4.2 Proportions analogiques à partir de couples de $sim_a = 0,75$

La première proportion analogiques à partir de couples de $sim_a = 0,75$ qui nous a intéressée mettait en jeu les lexies ACCOMPAGNATRICE **I** [*Jeanne est accompagnatrice pour les enfants voyageant sans leurs parents.*], ACCOMPAGNATRICE **II** [*Le chef de chœur sera assisté de son accompagnatrice musicale.*], ACCOMPAGNATEUR **I** [*Çà t'intéresse, un boulot d'accompagnateur dans une école ?*] et ACCOMPAGNATEUR **II** [*Il apprécie les morceaux où la voix est complétée d'un seul accompagnateur, un piano par exemple*].

| description de ACCOMPAGNATRICE **I** : description de ACCOMPAGNATRICE **II** |
|---|
| :: |
| description de ACCOMPAGNATEUR **I** : description de ACCOMPAGNATEUR **II** |

Le tableau 3.15 présente les caractéristiques communes des couples de descriptions en œuvre dans la première proportion analogique construite.

En observant ces couples, nous pouvons affirmer que « la description de ACCOMPAGNATRICE **I** est à la description de ACCOMPAGNATRICE **II** ce que la description de ACCOMPAGNATEUR **I** est à la description de ACCOMPAGNATEUR **II** », en ce sens qu'il y existe une similarité de Relations à propos des liens de FL décrits et que leurs Relations concernant les CG varient uniquement en terme de genre.

Nous avons choisi d'ajouter à ces observations les caractéristiques communes au couple de descriptions (ACCENTUATION **I.1** [*L'accentuation des syllables en anglais est importante.*],ACCENTUATION **I.2** [*L'accentuation des lettres majuscules est recommandée par l'Académie Française.*]) présenté dans le tableau 3.16.

À partir de ces caractéristiques, nous nous sommes intéressée à la proportion

| | ACCOMPAGNATRICE **I** : ACCOMPAGNATRICE **II** | ACCOMPAGNATEUR **I** : ACCOMPAGNATEUR **II** |
|---|---|---|
| Nbr CG | 2 | 2 |
| CG communes | `nom commun` fém. | `nom commun` masc. |
| Nbr FL en jeu | 2 | 2 |
| FL communes | **Syn**$_{\cap \text{ sexe}}$ **Fact$_2$** | **Syn**$_{\cap \text{ sexe}}$ **Fact$_2$** |
| Nbr liens de FL | 2 | 2 |
| Cibles communes | 0 | 0 |
| Vocable | = | = |

TAB. 3.15 : Comparaison de couples de descriptions de lexies de $sim_a = 0,75$

analogique suivante[21] :

description de ACCENTUATION **I.1** : description de ACCENTUATION **I.2**
::
description de ACCOMPAGNATEUR **I** : description de ACCOMPAGNATEUR **II**

Nous pouvons alors affirmer que « la description de ACCENTUATION **I.1** est à la description de ACCENTUATION **I.2** ce que la description de ACCOMPAGNATEUR **I** est à la description de ACCOMPAGNATEUR **II** ». En effet, si les Relations concernant les liens de FL ne sont pas identiques d'un couple à l'autre, nous sommes tout de même en présence d'une régularité qui s'observe dans les deux couples de façon analogue.

La première de ces deux proportions analogiques semble toutefois plus pertinente, en ce sens qu'elle permettrait de dégager une règle du type :

> Si les FL **Syn**$_{\cap \text{ sexe}}$ et **Fact$_2$** sont en jeu dans une DULB, alors elles le sont aussi dans une DULS du même vocable[22].

#### 3.5.4.3 Proportions analogiques à partir de couples de $sim_a \leq 0,20$

description de ENNEMI$_N$ **I** : description de ENNEMI$_N$ **III.a**
::
description de VOITURE **1** : description de VOITURE **2**

[21] Nous aurions pu tout aussi bien nous intéresser à la proportion ACCENTUATION **I.1** : ACCENTUATION **I.2** :: ACCOMPAGNATRICE **I** : ACCOMPAGNATRICE **II**.

[22] Il est bien entendu que cette seule proportion analogique ne permet pas d'extraire une telle règle et qu'il faudrait pour cela des données en nombre significatif. Il faut également s'attendre à ce que ces données amènent à prendre en considération la structure polysémique des vocables.

| | ACCENTUATION **I.1** : ACCENTUATION **I.2** |
|---|---|
| Nbr CG | 2 |
| CG communes | `nom commun` fém. |
| Nbr FL en jeu | 2 |
| FL communes | $\mathbf{V_0}$ $\mathbf{S_3}$ |
| Nbr liens de FL | 2 |
| Cibles communes | 0 |
| Vocable | = |

TAB. 3.16 : Caractéristiques communes au couple de descriptions (ACCENTUATION **I.1**, ACCENTUATION **I.2**)

Nous avons choisi de construire une proportion analogique à partir des couples de descriptions des lexies (ENNEMI$_N$ **I** [*Leurs relations avec le voisinage sont exécrables. Des ennemis partout.*], ENNEMI$_N$ **III.a** [*L'ennemi approchait. On entendait déjà les cris des cavaliers et les sabots des chevaux.*]) et (VOITURE **1** [*Ce pilote est aussi à l'aise au volant d'une voiture qu'au guidon d'une moto.*], VOITURE **2** [*Quand on prend le train pour Paris, on est souvent installé dans la voiture 18.*]), car il s'agissait de couples à faible $sim_a$ qui comprenaient des lexies dont la description nous semblait avancée au vu du nombre de liens de FL en présence.

Le tableau 3.17 présente les caractéristiques communes[23] de chacun de ces couples.

En observant ces caractéristiques, nous percevons une certaine analogie (CG communes, cible commune, vocable commun, variation du nombre et de la nature des liens de FL). Il semble cependant difficile d'établir des Relations pertinentes, tout autant que de faire émerger quelque règle intéressante que ce soit.

#### 3.5.4.4 Conclusion

À partir de ces observations, nous pouvons effectivement affirmer que les descriptions de lexies d'un même vocable ayant un certain degré de $sim_a$ entrent en relation analogique avec les descriptions de lexies de tout autre vocable ayant entre elles une $sim_a$ analogue.

Toutes les relations analogiques ainsi établies ne sont pas pour autant également pertinentes et il semble judicieux de rester concentré sur les couples de lexies

[23]Pour chacun des couples, il est possible en réalité de comptabiliser deux cibles communes de liens d'application de FL, mais nous n'avons reporté dans le tableau que celles qui appartenaient à la valeur d'application d'une même FL.

| | ENNEMI$_N$ **I** : ENNEMI$_N$ **III.a** | VOITURE **1** : VOITURE **2** |
|---|---|---|
| $sim_a$ | 0,22 | 0,2 |
| Nbr CG | 2 *vs* 3 | 2 |
| CG communes | `nom commun masc.` | `nom commun fém.` |
| Nbr FL en jeu | 15 *vs* 17 | 7 |
| FL communes | **Syn**$_\cap$, **A$_0$** **Magn**, **Oper$_1$** | **S$_1$** **S$_2$** |
| Nbr liens de FL | 18 *vs* 30 | 15 *vs* 21 |
| Cibles communes | 1 | 1 |
| Vocable | = | = |

TAB. 3.17 : Comparaison des couples de descriptions de $sim_a \leq 0,20$

partageant une $sim_a$ élevée.

### 3.5.5 Pertinence d'une distinction des unités lexicales de base

HYPOTHÈSE 4 : Les proportions analogiques permettant de faire émerger des règles pertinentes sont celles qui mettent en relation des DULB et des DULNB de la manière suivante :
$DULS_{voc1} : DULB_{voc1} :: DULS_{voc2} : DULB_{voc2}$

#### 3.5.5.1 Construction de proportions analogiques

Pour vérifier la validité de cette hypothèse, nous avons choisi de nous intéresser aux deux couples de descriptions de lexies $sim_a \geq 0,75$ n'appartenant pas au même vocable mis en évidence en 3.5.2. Il s'agit des descriptions des lexies INCONSCIENCE **II** et CONSCIENCE **II** d'une part, BOUCHER$_N$ **2** et BOULANGER$_N$ **2** d'autre part.

Aucune de ces descriptions n'est une DULB. Nous avons décidé de les mettre en relation avec leur DULB.

Nous avons alors construit manuellement les proportions analogiques suivantes, correspondant à notre hypothèse :

description de INCONSCIENCE **II** : description de INCONSCIENCE **I**
::
description de CONSCIENCE **II** : description de CONSCIENCE **I**

description de BOUCHER$_N$ **2** : description de BOUCHER$_N$ **1**
::
description de BOULANGER$_N$ **2** : description de BOULANGER$_N$ **1**

Nous avons, dans un premier temps, souhaité tester si ces proportions étaient plus pertinentes que leurs expressions équivalentes par permutation des moyens. Nous avons pour cela construit les deux proportions analogiques suivantes, comme points de comparaison :

description de INCONSCIENCE **II** : description de CONSCIENCE **II**
::
description de INCONSCIENCE **I** : description de CONSCIENCE **I**

description de BOUCHER$_N$ **2** : description de BOULANGER$_N$ **2**
::
description de BOUCHER$_N$ **1** : description de BOULANGER$_N$ **1**

Selon les conclusions de l'hypothèse précédente, ces proportions analogiques doivent être valides. Pour vérifier si celles mettant en œuvre le patron $DULS_{voc1}$ : $DULB_{voc1}$ :: $DULS_{voc2}$ : $DULB_{voc2}$ font émerger des règles plus pertinentes que les autres, nous nous sommes intéressée aux Relations similaires entre les couples de descriptions mises en relation analogique.

**A)** INCONSCIENCE et CONSCIENCE

Le tableau 3.18 présente les Relations que nous pouvons observer dans les proportions analogiques mettant en œuvre INCONSCIENCE **II** et CONSCIENCE **II**. Les valeurs de $sim_a$ qui y sont présentées, qui ne font l'objet d'aucune Relation, le sont à titre informatif.

Nous n'observons alors aucune différence en fonction de la proportion considérée, à l'exception d'une cible commune dans le cas des proportions correspondant à notre hypothèse. Cette cible commune, du lien d'application de la FL $\mathbf{A}_1$, était cependant une erreur et a été corrigée depuis. En effet, la cible du lien d'application de la FL $\mathbf{A}_1$ pour la lexie CONSCIENCE **II** était alors INCONSCIENT **II**.

L'organisation des proportions apparaît donc dans ce cas comme étant de peu d'importance.Aucune Relation pertinente ne se dégage, quelque soit les couples de lexies considérés.

**B)** BOUCHER$_N$ et BOULANGER$_N$

Le tableau 3.19 présente une comparaison des Relations que nous pouvons observer dans les proportions analogiques mettant en œuvre BOUCHER$_N$ **2** et BOULANGER$_N$ **2**. Contrairement au cas précédent, nous y observons des Relations différentes.

La régularité de variation de nombres et de natures de liens de FL que nous observons dans la proportion respectant le patron $DULS_{voc1}$ : $DULB_{voc1}$ :: $DULS_{voc2}$ :

| | (INCONSCIENCE II,CONSCIENCE II) *vs* (INCONSCIENCE I,CONSCIENCE I) | (INCONSCIENCE II,INCONSCIENCE I) *vs* (CONSCIENCE II,CONSCIENCE I) |
|---|---|---|
| $sim_a$ | 0,75 *vs* 0,67 | 0,75 |
| Nbr CG | 2 | 2 |
| CG communes | `nom commun`, `fém.` | `nom commun`, `fém.` |
| Nbr FL en jeu | 2 | 2 |
| FL communes | **Anti** et $\mathbf{A_1}$ | **Anti** et $\mathbf{A_1}$ |
| Nbr liens de FL | 2 | 2 |
| Cibles communes | 1 *vs* 0 | 0 |
| Vocable | $\neq$ | $=$ |

TAB. 3.18 : Comparaison des proportions analogiques incluant INCONSCIENCE II et CONSCIENCE II

$DULB_{voc2}$ semble être l'objet de Relations pertinentes.

Elle nous a permis d'énoncer les Relations similaires « **À un nombre de cinq liens d'application de FL dans une description de lexie A correspond un nombre de deux liens d'application de FL dans une description de lexie B** » et « **À la mise en jeu des FL $\mathbf{S_3}$, $\mathbf{Syn_{\cap\ sexe}}$, $\mathbf{type\ particulier\ S_2}$ et $\mathbf{S_{loc}\&S_4}$ dans une description de lexie A, correspond la mise en jeu des FL $\mathbf{Syn_{\cap\ sexe}}$ et $\mathbf{S_{2\cap}}$ dans une description de lexie B** ».

À partir de telles Relations, nous pensons pouvoir extraire des règles de type :

> Si une FL $\mathbf{Syn_{\cap\ sexe}}$ est en jeu dans une DULS, elle l'est aussi dans celle de la DULB du même vocable.
>
> Si la FL $\mathbf{type\ particulier\ S_2}$ est en jeu dans une DULS, la FL $\mathbf{S_{2\cap}}$ l'est dans la DULB du même vocable.
>
> Si les FL $\mathbf{Syn_{\cap\ sexe}}$, $\mathbf{type\ particulier\ S_2}$, $\mathbf{S_3}$ et $\mathbf{S_{loc}\&S_4}$ sont en jeu dans une DULS, les FL $\mathbf{Syn_{\cap\ sexe}}$ et $\mathbf{S_{2\cap}}$ le sont dans la DULB du même vocable.

Bien entendu, un unique cas de proportions analogiques ne permet pas d'établir de telles règles.

Dans un second temps, nous avons souhaité comparer la proportion analogique respectant le patron $DULS_{voc1} : DULB_{voc1} :: DULS_{voc2} : DULB_{voc2}$ à la proportion analogique non équivalente suivante :

| description de $\text{BOUCHER}_\text{N}$ **2** : description de $\text{BOUCHER}_\text{N}$ **1** |
|---|
| :: |
| description de $\text{BOULANGER}_\text{N}$ **1** : description de $\text{BOULANGER}_\text{N}$ **2** |

| | (BOUCHER$_N$ **2**,BOULANGER$_N$ **2**) *vs* (BOUCHER$_N$ **1**,BOULANGER$_N$ **1**) | | (BOUCHER$_N$ **2**,BOUCHER$_N$ **1**) *vs* (BOULANGER$_N$ **2**,BOULANGER$_N$ **1**) |
|---|---|---|---|
| $sim_a$ | 0,75 | 0,67 | 0,440 |
| Nbr CG | 2 | | 2 |
| CG communes | `nom commun`, `masc.` | | `nom commun`, `masc.` |
| Nbr FL en jeu | 5 | 2 | 5→2 |
| FL communes | $\mathbf{S_3}$, $\mathbf{Syn_{\cap\ sexe}}$ $\mathbf{S_{type\ particulier}}$ $\mathbf{S_2}$ $\mathbf{S_{loc}\&S_4}$ | $\mathbf{Syn_{\cap\ sexe}}$ $\mathbf{S_{2\cap}}$ | $\mathbf{Syn_{\cap\ sexe}}$ |
| Nbr liens de FL | 4 | 2 | 4→2 |
| Cibles communes | 2 | 0 | 0 |
| Vocable | ≠ | | = |

TAB. 3.19 : Comparaison des proportions analogiques équivalentes mettant en jeu BOUCHER$_N$ **2** et BOULANGER$_N$ **2**

Le tableau 3.20 présente une comparaison des Relations que nous pouvons observer dans les deux proportions analogiques non équivalentes mettant en œuvre BOUCHER$_N$ **2** et BOULANGER$_N$ **2**.

Nous constatons qu'il est bien plus difficile d'établir des Relations pertinentes à partir de la dernière proportion créée. En effet, l'alternance en nombre de FL en jeu et en nombre de liens de FL n'est plus régulière.

Seule la règle suivante peut être énoncée à partir des Relations observées :

> Si une FL $\mathbf{Syn_{\cap\ sexe}}$ est en jeu dans la description d'une lexie, elle l'est aussi dans celle d'une autre lexie appartenant au même vocable.

#### 3.5.5.2 Conclusion

Si l'ordre des lexies dans la proportion analogique s'est avéré n'être d'aucune pertinence pour la mise en relation des lexies des vocables CONSCIENCE et INCONSCIENCE, il a permis en revanche de faire émerger des règles de la mise en relation des lexies des vocables BOUCHER et BOULANGER.

### 3.5.6 Filtrage assisté par le score de Turney

> HYPOTHÈSE 5 : Le calcul de score proposé par Turney (2006) peut être utilisé comme filtre pour se concentrer sur les analogies pertinentes.

| | (BOUCHER$_N$ **2**,BOUCHER$_N$ **1**) *vs* (BOULANGER$_N$ **2**,BOULANGER$_N$ **1**) | (BOUCHER$_N$ **2**,BOUCHER$_N$ **1**) *vs* (BOULANGER$_N$ **1**,BOULANGER$_N$ **2**) |
|---|---|---|
| $sim_a$ | 0,44 | 0,44 |
| Nbr CG | 2 | 2 |
| CG communes | `nom commun`, `masc.` | `nom commun`, `masc.` |
| Nbr FL en jeu | 5→2 | 5→2 *vs* 2→5 |
| FL communes | **Syn$_{\cap \text{ sexe}}$** | **Syn$_{\cap \text{ sexe}}$** |
| Nbr liens de FL | 4→2 | 4→2 *vs* 2→4 |
| Cibles communes | 0 | 0 |
| Vocable | = | = |

TAB. 3.20 : Comparaison des proportions analogiques non équivalentes mettant en jeu BOUCHER$_N$ **2** et BOULANGER$_N$ **2**

Dans les sections 3.5.4 et 3.5.5, nous avons observé des proportions analogiques, dont nous avons estimé que certaines étaient plus pertinentes que d'autres.

Nous avons donc choisi de regarder le *degré d'analogicité* qui leur serait attribué par application du score de Turney (2006), pour tester notre dernière hypothèse[24].

Le tableau 3.21 montre les résultats obtenus pour les trois proportions mises en avant dans la section 3.5.4. Nous pouvons y voir que celle qui nous paraissait la plus pertinente est la plus proche, avec un score de $0,667$, tandis que celle qui nous semblait la moins pertinente est la plus lointaine, avec un score de $0,085$.

Le tableau 3.22 montre les résultats obtenus pour les cinq proportions mises en avant dans la section 3.5.5. Nous pouvons y voir que les deux qui nous paraissaient équivalentes sont presque également *proches*, avec des scores de $0,7085$ et $0,75$. Entre les trois restantes, celle organisant la proportion analogique selon le modèle $DULS : DULB :: DULS : DULB$ entre descriptions de lexies de mêmes vocables est la plus *proche*, avec un score de $0,7085$, contre $0,424$ et $0,375$.

#### 3.5.6.1 Conclusion

L'utilisation du score de Turney pour filtrer les relations analogiques paraît donc envisageable. Il nécessite cependant d'établir un seuil de filtrage.

Une rapide observation des proportions analogiques composées de couples $DULS : DULB$ appartenant à un même vocable et partageant une $sim_a \geq 0,75$ d'une part, et comprise entre $0,5$ et $0,74$ d'autre part, nous a permis d'estimer ce seuil à $0,4$.

[24] L'ordre des proportions de la section 3.5.4 a été réorganisé entre DULB et DULS lorsque cela était possible.

| proportion analogique $A : B :: C : D$ | $sim_a(A, C)$ | $sim_a(B, D)$ | $1/2(sim_a(A, C) + sim_a(B, D))$ |
|---|---|---|---|
| ACCOMPAGNATRICE **II** : ACCOMPAGNATRICE **I** :: ACCOMPAGNATEUR **II** : ACCOMPAGNATEUR **I** | 0,667 | 0,667 | 0,667 |
| ACCENTUATION **I.2** : ACCENTUATION **I.1** :: ACCOMPAGNATEUR **II** : ACCOMPAGNATEUR **I** | 0,18 | 0,25 | 0,215 |
| ENNEMI$_N$ **III.a** : ENNEMI$_N$ **I** :: VOITURE **2** : VOITURE **1** | 0,09 | 0,07 | 0,085 |

TAB. 3.21 : Degré d'analogicité des proportions analogiques de la section 3.5.4

| proportion analogique $A : B :: C : D$ | $sim_a(A, C)$ | $sim_a(B, D)$ | $1/2(sim_a(A, C) + sim_a(B, D)$ |
|---|---|---|---|
| INCONSCIENCE **II** : CONSCIENCE **II** :: INCONSCIENCE **I** : CONSCIENCE **I** | 0,75 | 0,75 | 0,75 |
| INCONSCIENCE **II** : INCONSCIENCE **I** :: CONSCIENCE **II** : CONSCIENCE **I** | 0,75 | 0,667 | 0,7085 |
| BOUCHER$_N$ **2** : BOULANGER$_N$ **2** :: BOUCHER$_N$ **1** : BOULANGER$_N$ **1** | 0,44 | 0,44 | 0,44 |
| BOUCHER$_N$ **2** : BOUCHER$_N$ **1** :: BOULANGER$_N$ **2** : BOULANGER$_N$ **1** | 0,75 | 0,667 | 0,7085 |
| BOUCHER$_N$ **2** : BOUCHER$_N$ **1** :: BOULANGER$_N$ **1** : BOULANGER$_N$ **2** | 0,375 | 0,375 | 0,375 |

TAB. 3.22 : Degré d'analogicité des proportions analogiques de la section 3.5.5

Ces observations demandent cependant à être approfondies et élargies à l'ensemble des couples ayant une $sim_a \geq 0,5$.

## Conclusions et perspectives

Les données du RL-fr utilisées pour réaliser cette première expérience étaient insuffisantes pour valider ou invalider les hypothèses énoncées. Les analyses menées nous ont cependant permis de nous conforter dans l'idée qu'il est possible de tester de telles hypothèses sur les données et nous pouvons en résumer ici les conclusions.

La détection de relations analogiques pertinentes entre descriptions de lexies, susceptibles de faire émerger des règles lexicales, semble possible. Certains cas prometteurs ont été observés et l'existence de Relations communes entre lexies de $sim_a$ élevée, aussi bien deux à deux que toutes regroupées, a été esquissée.

L'utilisation d'une mesure de distance d'édition canonique a bien fonctionné, même si un certain nombre de cas rencontrés, non évoqués dans ce chapitre, mettaient en avant un problème d'organisation des Attributs. Les descriptions des lexies CONTEUR **1** [*Noam aime quand son père lui lit des histoires. Il trouve que c'est un bon conteur.*] et CONTEUR **2** [*Tu préfères quel conteur, toi, Andersen ou les frères Grimm ?*], par exemple, ont été considérées comme moins similaires qu'elles n'auraient dû. Si leurs descriptions mettaient bien en jeu les deux mêmes FL, leurs positions étaient inversées. Pour remédier à ce problème, nous aurions dû contraindre davantage l'ordre de sélection des éléments de description dans la base de données et leur transcription dans le format RDF N3. Des informations de position des FL sont disponibles dans la base[25]. Nous aurions pu utiliser ces informations, couplées à un tri sur les identifiants numériques à cette fin.

Pour cette première expérience, nous avions mis à plat les descriptions des lexies du RL-fr. Cette mise à plat ne nous permet pas de graduer l'importance des différentes caractéristiques des lexies. Elle ne nous permet pas non plus de distinguer les éléments de description relationnels, comme le sont les liens de FL, de ceux qui ne le sont pas, comme les CG. Enfin, elle ne permet aucune distinction entre liens de FL syntagmatiques et paradigmatiques. Une prise en compte de cette différence serait pourtant à rapprocher d'une distinction entre la similarité et la contiguïté, présentées en 3.1 — à la suite de Jakobson (1963) — comme le fondement de la sélection et de la combinaison.

L'organisation des proportions analogiques entre DULB et DULS fondée sur cette mesure s'est toutefois avérée prometteuse. Nous pouvons donc envisager de mettre au point une méthode de détection de relations analogiques entre couples de descriptions de lexies de même vocable. Une telle application permettrait de faire émerger des règles lexicales utiles à la semi-automatisation de l'activité lexicographique. Dans le cas des vocables BOULANGER et BOUCHER, nous avons pu observer des régularités qui sont encourageantes pour la réalisation de patrons polysémiques, tels que ceux mis en évidence dans les travaux de Barque (2008).

Nous avons commencé à approfondir cette question dans le cadre d'une seconde expérience. Nous nous sommes intéressée aux vocables polysémiques dont la description était avancée. L'objectif était alors de visualiser la similarité entre les des-

[25] Ces informations sont utilisées pour la visualisation des FL dans l'éditeur lexicographique.

criptions des lexies à l'intérieur de chaque vocable.

Nous nous sommes servie pour cela de la même méthode de comparaison que dans notre première expérience. La visualisation a été effectuée à l'aide de la librairie JavaScript D3[26]. La figure 3.4 présente un exemple d'une telle représentation. Dans cet exemple, la similarité entre descriptions a été calculée à partir des éléments suivants : identifiant du vocable, identifiant de la lexie, numérotation, CG, FL et étiquette sémantique quand elle était disponible.

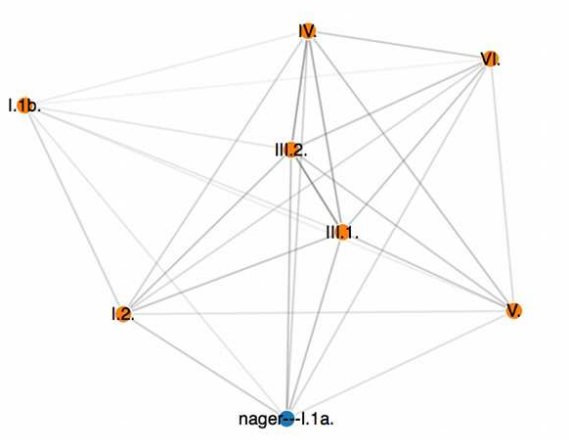



FIG. 3.4 : Visualisation de la similarité entre descriptions du vocable NAGER

Nous avons également effectué quelques essais fondés uniquement sur les liens de FL, en dissociant ou non les liens syntagmatiques et paradigmatiques. La figure 3.5 montre comment l'aspect général du vocable NAGER varie en fonction des éléments de descriptions comparés. Elle présente, de gauche à droite, une visualisation de la structure polysémique de NAGER établie à partir de tous les liens de FL, puis des liens de FL syntagmatiques seuls et enfin des liens de FL paradigmatiques. Les deux premières visualisations sont très proches l'une de l'autre. Elles ne sont pas très éloignées non plus de celle présentée dans la figure 3.4. La dernière, en revanche, est radicalement différente.

FIG. 3.5 : Variations de la similarité entre descriptions du vocable NAGER en fonction des éléments comparés

À la suite de ces deux expériences, nous pensons que la mise au point d'une méthode de détection de relations analogiques entre couples de descriptions de lexies de

[26]Datat Driven Documents `http://d3js.org/`

même vocable nécessite la prise en compte des différents types de relation de copolysémie entre ces lexies. Les versions plus récentes du RL-fr encodent ces relations. Elles pourraient donc être utilisées pour réaliser de nouvelles expériences. Ce serait notamment l'occasion de s'intéresser à la distinction entre similarité et contiguïté, que Jakobson (1963) rapporte l'une à la métaphore, l'autre à la métonymie.

Notre première expérience nous a également amenée à considérer comme intéressante la détection de relations analogiques au-delà des couples de lexies de même vocable. Cette seconde question a, elle aussi, fait l'objet d'une expérience supplémentaire. Nous nous sommes intéressée aux descriptions des unités lexicales comportant l'étiquette sémantique sport. L'objectif de cette expérience était de tester s'il était envisageable d'utiliser notre méthode pour obtenir un prototype de lexie associée à une étiquette sémantique donnée. La faible quantité de données disponible ne nous permettait pas de répondre clairement à cette question. Nous avons tout de même observé que notre mesure retournait des résultats intéressants. En effet, sur les onze lexies comparées, notre méthode a automatiquement rapproché les quatre descriptions de sport collectif se jouant avec une balle que sont FOOTBALL **1**, RUGBY **1**, VOLLEY-BALL **1** et $\text{BASKET}_{\text{N, masc.}}$ **1**. Les descriptions de ces lexies sont apparues comme étant les plus similaires. Nous avons également observé que la construction de proportions analogiques à partir de couples composés de ces quatre DULB et des DULS de leur vocable respectif fonctionnait bien.

La dernière chose que nous apprend notre expérience initiale, c'est que le filtrage des relations analogiques détectées semble indispensable. Le degré d'analogicité testé ici, fondé sur le calcul de score de Turney (2006), fournit des résultats proches de notre intuition. Dans la suite de nos travaux, son utilisation sera toutefois couplée à une validation humaine, par l'intermédiaire des lexicographes du RL-fr.

# Chapitre 4

# Composantes connexes analogues

## Sommaire

# Introduction

Les résultats de l'expérience initiale présentée dans le chapitre 3 sont encourageants. Plusieurs types d'exploration du RL-fr par raisonnement analogique semblent envisageables. Les trois points d'entrée dans le réseau que nous avons testés — par sélection des descriptions de lexies nominales ayant au minimum deux liens de FL sortants et appartenant à des vocables polysémiques, par étiquette sémantique et par état d'avancement du travail lexicographique — couplés aux différents éléments d'informations comparés — CG, appartenance à un vocable, ensembles des liens de FL et de leurs cibles, étiquettes sémantiques, distinction des liens de FL syntagmatiques et paradigmatiques — peuvent être vus comme des dimensions distinctes, mettant en évidence différents types de rapports.

Comme nous l'avons souligné, la mise à plat des descriptions des lexies n'est cependant pas satisfaisante. Tandis que certains éléments du graphe lexical qu'elles constituent sont de nature relationnelle, d'autres paraissent davantage de l'ordre de l'Attribut. L'expérience détaillée dans ce chapitre prend en compte cette distinction[1]. Elle a été réalisée à la suite d'une réflexion sur les propriétés formelles du RL-fr, détaillée dans le chapitre 2.

Cette analyse formelle a attiré notre attention sur l'organisation du RL-fr en zones denses. Formées de lexies fortement interconnectées par l'intermédiaire de relations lexicales, ces zones denses forment des agrégats lexicaux. Nous pensons que la nature de ces agrégats varie en fonction de leur taille. Tandis que les agrégats de grandes tailles correspondraient à des champs sémantiques, les agrégats plus denses et plus petits correspondraient à des connexions lexicales particulières.

Nous avons choisi de concentrer notre attention sur ce second type d'agrégats. Ce chapitre présente la première expérience que nous leur avons consacrée. Il commence par introduire nos hypothèses, à travers la notion de configurations de dérivations lexicales. Il détaille ensuite notre réflexion sur la sélection des données et la circonscription d'une première étape : le regroupement de sous-graphes analogues. Tout comme le chapitre 3, il se poursuit par une présentation de la mise en place technique de l'expérience et des résultats obtenus.

## 4.1 Configurations de dérivations lexicales

Tout au long du premier chapitre, nous avons utilisé l'exemple d'une analogie perçue par les locuteurs du français entre les vocables ABOYER et BEUGLER. Nous avons petit à petit développé l'idée que cette analogie n'était pas basée sur une seule proportion analogique mettant en œuvre quatre unités lexicales, mais sur un ensemble interdépendant plus large. La figure 4.1 propose deux sous-graphes lexicaux[2] permettant de rendre compte d'une telle analogie.

[1] Une partie du matériau inclus dans ce chapitre a été publiée dans Ollinger (2014).

[2] Ces sous-graphes sont inspirés du RL-fr, mais n'en sont pas directement extraits. La numérotation des lexies qu'ils comportent est notamment arbitraire.

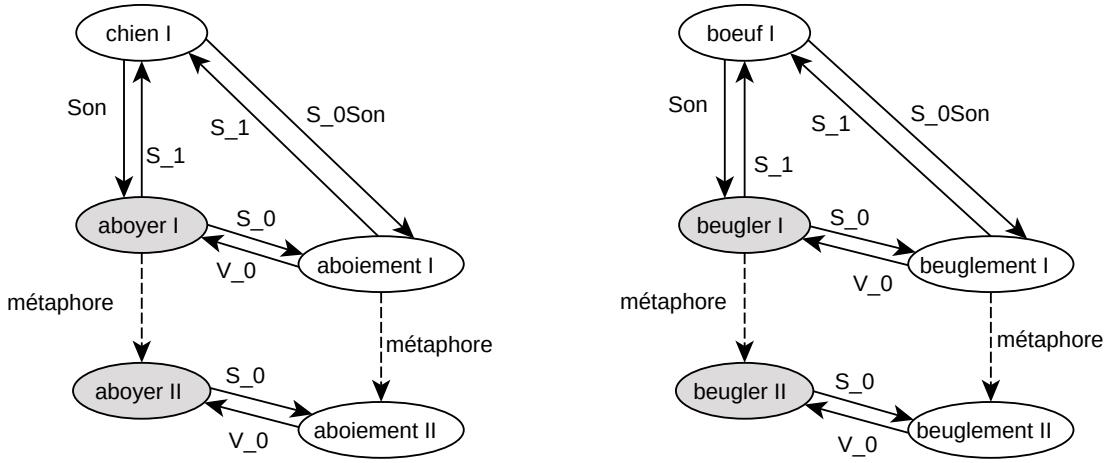



FIG. 4.1 : Similarité de connexions lexicales

L'observation de ces deux sous-graphes permet de constater qu'ils mettent en jeu des relations lexicales identiques entre des sommets lexicaux distincts. Par ailleurs, nous pouvons supposer que les descriptions encapsulées dans ces sommets partagent une certaine similarité deux à deux. Ainsi, ABOIEMENT I et BEUGLEMENT I disposent certainement tous deux de la CG `nom commun`, de l'étiquette sémantique cri d'animal et d'une même instance de FP prenant respectivement la forme `aboiement de X` et `beuglement de X`.

Nous pouvons aisément imaginer qu'une telle corrélation entre relations lexicales et éléments de description ne soit pas un fait isolé. Qu'il existe, d'une part, d'autres ensembles de lexies analogues à ces deux-là et, d'autre part, d'autres ensembles de lexies relevant de *connexions lexicales récurrentes entre lexies particulières* se répétant à l'intérieur du graphe.

Étant donné la topologie du RL-fr, nous pensons qu'une exploration par raisonnement analogique des agrégats lexicaux de petite taille permettra de mettre en évidence ces connexions particulières et qu'il sera alors possible d'en élaborer des modèles.

De tels modèles, que nous nommons *configurations de dérivations lexicales*, pourraient être exploités pour enrichir le RL-fr et intégrer de nouvelles fonctionnalités à l'éditeur lexicographique utilisé pour son développement. Ils seraient constitués d'un ensemble de relations orientées entre lexies, ou plus exactement entre profils de lexies détaillant les caractéristiques nécessaires au déclenchement d'une configuration. Deux axes d'enrichissement automatique seraient alors envisageables : la génération de liens entre lexies correspondant aux profils et l'enrichissement des descriptions incomplètes de lexies d'ores et déjà interconnectées.

Le cas des sous-graphes de la figure 4.1 pourrait ainsi aboutir à la configuration présentée dans la figure 4.2. Les sommets lexicaux y sont remplacés par l'ensemble des éléments communs des descriptions qu'ils encapsulent. Les liens de FL sortants sont également fournis, regroupés par familles afin de limiter la contrainte. Celles

permettant d'encoder la synonymie, l'antonymie et les expressions contrastives sont exclues.

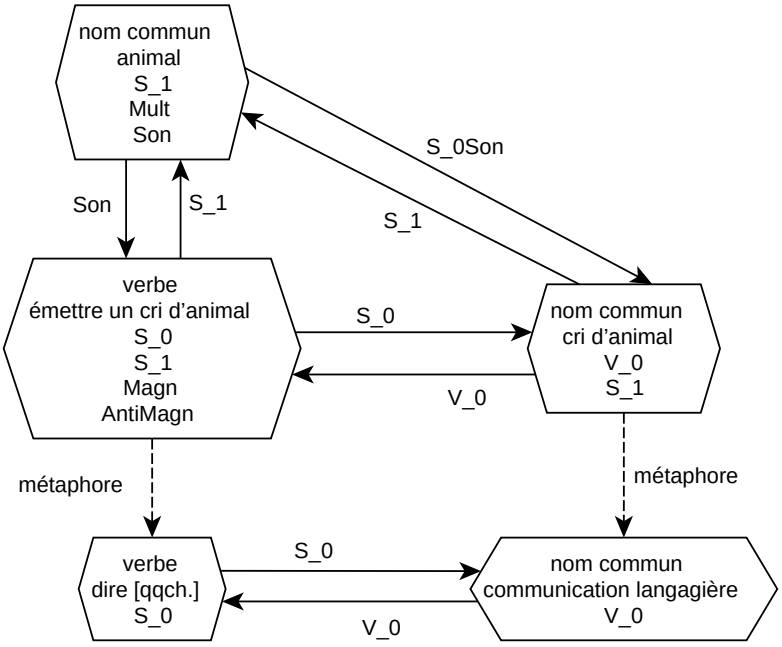



FIG. 4.2 : Exemples de configuration de dérivations lexicales

Ainsi, la comparaison des lexies CHIEN **I** et BŒUF **I** permet d'élaborer, par abstraction, le profil <nom commun, animal, **S₁**, **Mult**, **Son**>. Un sommet lexical quelconque ne peut donc occuper leur place dans la configuration que s'il s'agit d'un nom commun, ayant pour étiquette sémantique animal et étant la source de liens de FL ayant pour cible au moins une nominalisation de son premier actant sémantique, un collectif et une lexie dénotant un son[3].

L'ensemble des profils proposés ici comporte des caractéristiques grammaticales, des étiquettes sémantiques et des familles de FL. Il est cependant possible d'envisager que certains de ces éléments soient absents ou que d'autres viennent s'y ajouter, comme des FP.

## 4.2 Délimitation de sous-graphes

L'exploration par raisonnement analogique des agrégats lexicaux de petite taille du RL-fr peut être réalisée à partir du dénombrement des sous-graphes qui le composent.

[3]Il est important de noter que certains des élements de ce profil pourraient être remis en cause par l'ajout d'un troisième sous-graphe mettant en jeu les mêmes connexions lexicales. Ainsi, si les descriptions des noms d'animaux tendent à contenir des FL de la famille **Mult**, nous pouvons d'ores et déjà prévoir que cet élément disparaîtrait si la lexie CHAT[1] **I.a** entrait en jeu dans une telle configuration de connexions lexicales.

Diverses études ont été menées sur les types de sous-graphes existants et l'information qu'ils véhiculent. Borgatti et al. (1990), qui s'intéressent à la notion de cohésion dans les réseaux sociaux, soulignent l'importance de la prise en compte du sujet d'étude dans le choix d'un modèle de sous-graphes à exploiter.

### 4.2.1 Cliques

Nous avons évoqué, au chapitre 2, les travaux de Ploux et Victorri (1998), qui proposent un traitement de la polysémie par délimitations d'espaces sémantiques à l'aide de dictionnaires de synonymes. Leur exploration se base sur la notion de *clique*, définit comme étant « un ensemble le plus grand possible de sommets [...] tous reliés deux à deux ». Le sous-graphe que nous avons construit autour du vocable ABOYER dans la figure 4.1 ne répond pas à cette définition, car les sommets CHIEN I, ABOYER II et ABOIEMENT II ne sont pas reliés. Il comporte deux cliques[4], présentées de manière dissociée dans la figure 4.3.

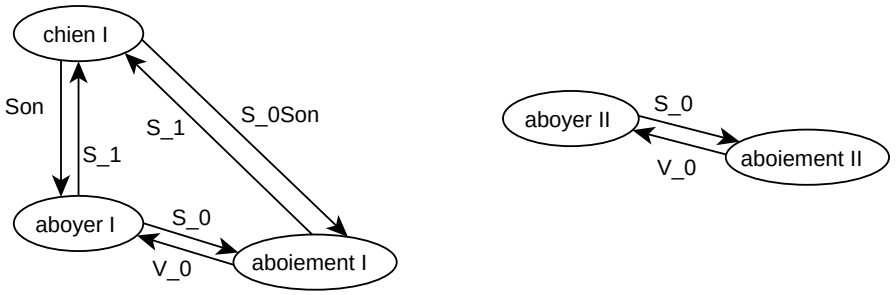



FIG. 4.3 : Clique lexicale

Bien que la première des deux cliques forme un ensemble de connexions lexicales pour lequel il semble intéressant d'établir un modèle, la seconde correspond simplement à un cas connu de FL inverses. De plus, ce découpage écarte la possibilité de s'intéresser à la coexistence d'une symétrie de dérivations syntaxiques entre les couples de lexies (ABOYER I, ABOIEMENT I) et (ABOYER II, ABOIEMENT II), d'une part, et d'une symétrie de dérivation métaphorique entre les couples (ABOYER I, ABOYER II) et (ABOIEMENT I, ABOIEMENT II), d'autre part. Leur exploration ne répond donc pas à notre besoin.

### 4.2.2 Communautés

Borgatti et al. (1990), Navarro et al. (2010) et Navarro (2013), pour leur part, concentrent leur attention sur la détection et l'analyse de *communautés*. De tels sous-graphes correspondent à des zones dont les sommets sont davantage connectés entre eux qu'au reste du graphe. Le sous-graphe que nous avons construit autour du vocable ABOYER ne répond pas non plus à cette définition. Si nous observons, à

[4]Si nous ne comptabilisons ici que deux cliques, c'est que nous nous appuyons sur la définition de Ploux et Victorri (1998), qui, à la suite de Bron et Kerbosch (1973), considèrent qu'une clique ne peut être contenue dans aucune autre clique.

l'aide de la figure 4.4, la connectivité du seul sommet ABOYER I dans son état actuel[5] dans le RL-fr, nous constatons qu'il est davantage connecté au reste du graphe qu'aux sommets qui nous intéressent, en gris.

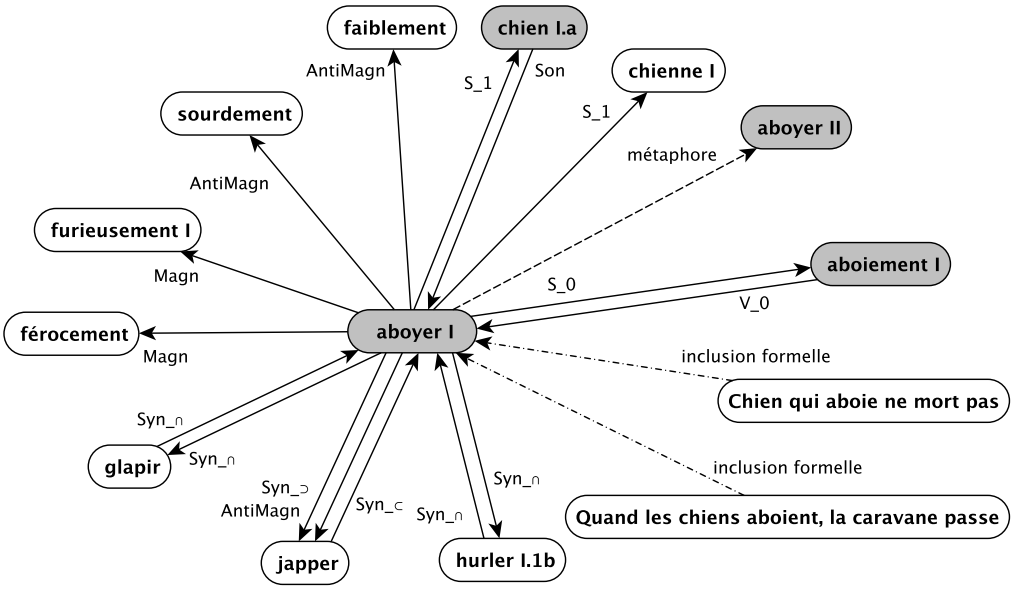



FIG. 4.4 : Connectivité du sommet ABOYER I

Nous pouvons cependant observer que si nous nous concentrons sur les seuls liens dérivationnels de cette figure (**$S_1$**, **$S_0$**, **$V_0$**, **Son**, métaphore), le sommet ABOYER I est davantage connecté aux sommets gris qu'au reste du graphe.

Sans effectuer de telle préselection d'arcs, nous émettons l'hypothèse que les communautés du RL-fr correspondent à des agrégats de grandes tailles, dont nous avons d'ores et déjà supposé qu'ils relevaient de champs sémantiques. Leur exploration ne répond pas à notre besoin immédiat et nous ne nous y intéresserons pas dans le cadre de nos expériences.

### 4.2.3 Motifs

Un troisième type de sous-graphes semble davantage correspondre à nos besoins. Il s'agit des *motifs locaux*. Tabourier (2010) définit ces motifs comme des « sous-graphes mettant en jeu un « petit » nombre de nœuds (typiquement $\leq 5$) – la taille du motif ». Il évoque leur exploitation dans le cadre de travaux sur les réseaux sociaux ainsi qu'en biologie et en génétique, parmi lesquels ceux d'Albert et Albert (2004), qui développent une méthode d'identification de patrons d'interactions entre protéines. Milo et al. (2002) et Milo et al. (2004) s'intéressent également à ces motifs. Dans la cadre d'une classification de graphes petit monde, ils comparent leurs sous-graphes de taille 3 et 4 à ceux de graphes aléatoires afin de faire émerger les plus spécifiques. Ces motifs spécifiques sont alors considérés comme une caractéristique permettant de regrouper les graphes petit monde en « superfamilles » partageant la

[5]Cette connectivité a été obtenue à l'aide d'un outil de visualisation que nous avons développé pour l'équipe de lexicographes. Nous avons interrogé cet outil le 23 septembre 2014.

même structure locale.

Nous ferons, pour notre part, la distinction entre les sous-graphes, que nous considérons comme des *occurrences de motif* et les *motifs* eux-mêmes, correspondant à un nombre et une répartition particulière d'arcs pour une taille donnée. Pour un graphe orienté simple, nous considérons donc que treize motifs de taille 3 existent, quel que soit le nombre d'occurrences de chacun d'entre eux. La figure 4.5 présente ces motifs. Pour un multigraphe orienté comme le RL-fr, les possibilités de nombre et d'organisation des arcs augmentent, ainsi que le nombre de motifs de chaque taille.

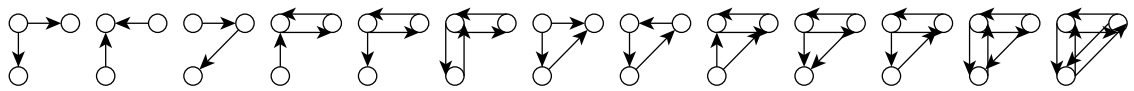


FIG. 4.5 : Motifs de taille 3 d'un graphe simple orienté

Le sous-graphe que nous avons construit autour du vocable ABOYER est une occurrence de motif de taille 5. Il comporte lui-même plusieurs occurrences de motifs de plus petite taille. La figure 4.6 fournit un exemple d'occurrence de chacun des deux motifs de taille 2 en présence. Le motif de gauche apparaît quatre fois — mettant en jeu des sommets et des arcs différents — celui de droite deux fois. Comme nous l'avons évoqué pour les cliques, l'observation de sous-graphes à deux sommets ne nous apprend rien de nouveau sur les connexions lexicales du RL-fr.

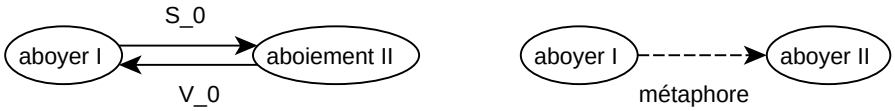



FIG. 4.6 : Exemples d'occurrences de motifs de taille 2

La figure 4.7, page suivante, fournit un exemple d'occurrence de chacun des trois motifs de taille 3 en présence. Le motif en haut à gauche apparaît deux fois, celui en-dessous quatre et celui de droite une seule. Ce dernier est particulièrement intéressant : il correspond à la clique pour laquelle nous avons souligné l'intérêt d'établir un modèle.

Pour sa part, la figure 4.8 qui l'accompagne, fournit un exemple d'occurrence de chacun des trois motifs de taille 4 en présence. Le motif du haut apparaît deux fois, celui en bas à gauche une seule et celui en bas à droite deux. Les deux premiers exemples semblent particulièrement intéressants. Si les connexions lexicales qu'ils comportent s'avèrent être récurrentes dans le RL-fr, ils pourraient donner lieu à des modèles exploitables dans le cadre de travaux sur la polysémie régulière, tels que ceux menés par Barque (2008), Barque et Chaumartin (2009), Goossens (2011) et Sikora (2014).

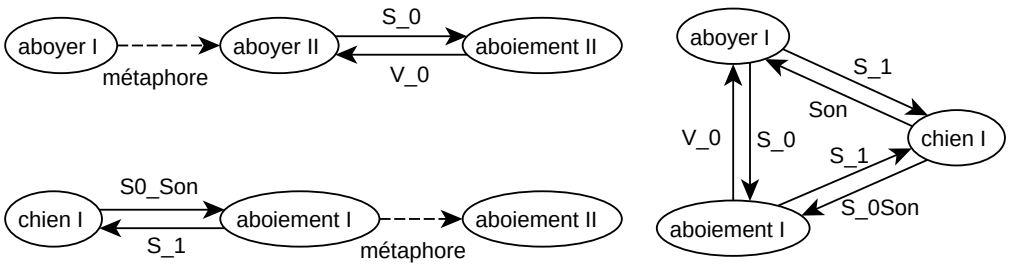



FIG. 4.7 : Exemples d'occurrences de motifs de taille 3

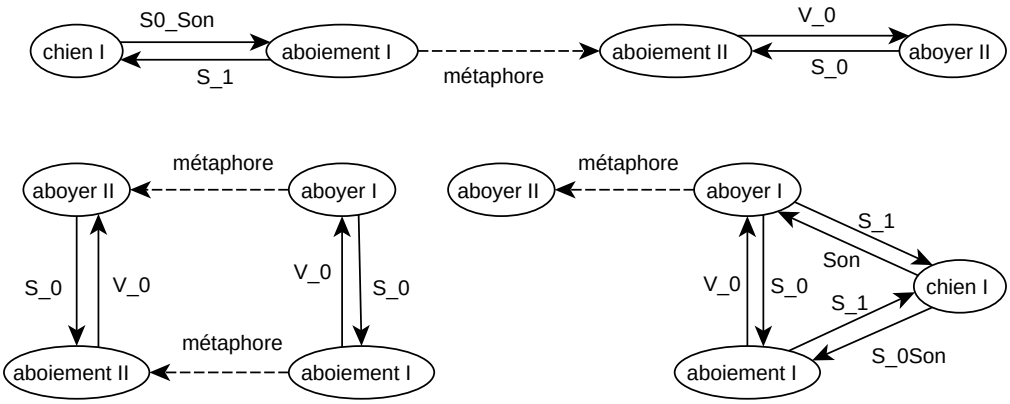



FIG. 4.8 : Exemples d'occurrences de motifs de taille 4

Le dénombrement des motifs d'un graphe n'est pas un problème algorithmique trivial. Les outils actuellement disponibles utilisent des méthodes gourmandes en temps d'exécution et en mémoire. Belderrar (2011) et Ribeiro (2011) proposent tous deux un panorama des algorithmes associés à ces méthodes. Tous ne sont pas adaptés au cas des multigraphes orientés. De plus, certaines de leurs implémentations ne retournent pas les occurrences de motifs, mais des listes de fréquences et de spécificités pour chacun des motifs possibles.

Dans le cadre de nos expériences, nous avons choisi d'utiliser les fonctionnalités dédiées aux motifs de la librairie python *graph-tool*[6] développée par Peixoto (2014). Elles implémentent l'algorithme proposé par Wernicke (2006).

Ces expériences, comme l'analyse topologique du RL-fr présentée au chapitre 2 et l'ensemble des manipulations qui seront dorénavant évoquées, ont été réalisées sur des versions du RL-fr exportées de la base de données au format GraphML (Brandes et al., 2002).

Le temps d'exécution de l'algorithme de Wernicke dépend du nombre de sommets et d'arcs du graphe explorés, ainsi que du degré moyen de ses sommets, du nombre de motifs qu'il comporte et du nombre d'occurrences de chacun d'entre eux. Wernicke et Rasche (2006) annoncent un temps d'exécution de 10 secondes pour le

[6] `http ://graph-tool.skewed.de/`

dénombrement et la mesure de spécificité des motifs de taille 5 du réseau *E.COLI* de Shen-Orr et al. (2002). Il s'agit d'un graphe orienté[7] de 423 sommets et 519 arcs. Le degré moyen de ses sommets est de 1,2. Il comporte 1 433 502 sous-graphes de taille 5, correspondant à 83 motifs distincts.

Comme nous l'avons vu au chapitre 2, le RL-fr est un graphe bien plus étendu et complexe que *E.COLI*. Lors de nos expériences, le dénombrement de ses motifs et la collecte de leurs occurrences ont donc nécessité plusieurs heures d'exécution. Comme nous l'avons évoqué au chapitre 1, l'exploration automatique par raisonnement analogique pose également d'importants problèmes d'optimisation. Ces considérations informatiques nous ont amenée à dissocier les tâches de mise en place d'une procédure d'identification de configurations de dérivations lexicales par regroupement de microstructures analogues et d'application de cette procédure aux occurrences de motifs. Nous avons alors choisi de débuter nos expériences en exploitant les sous-graphes plus facilement accessibles que sont les composantes connexes.

### 4.2.4 Composantes connexes

Nous l'avons vu au cours du chapitre 2, le RL-fr n'est pas un graphe connexe dans son état actuel de développement. Le 14 août 2014, il était possible de le partitionner en 12 987 composantes fortement connexes ou en 2 491 composantes faiblement connexes.

Nous avons émis l'hypothèse que les plus petites de ces composantes correspondaient à des paradigmes sémantiques et qu'ils étaient liés à la création de lexies à la volée, comme cible de liens de FL et participant à la nomenclature indirectement induite que nous avons présenté au chapitre 2, section 2.1.3. Nous avons également supposé qu'un travail en profondeur d'au moins une lexie de ces paradigmes aboutirait à son intégration à une composante de plus grande taille, à l'aide de liens de copolysémie et de FL syntagmatiques.

Contrairement aux occurrences de motifs, les composantes connexes, en tant que résultats d'une partition, comportent toutes des sommets distincts. Leur exploration limite donc les possibilités d'établir des configurations de dérivations lexicales, en limitant la participation de chaque lexie à un seul ensemble de connexions lexicales, le plus grand possible pour cette lexie.

La question s'est naturellement posée de savoir s'il était préférable d'explorer les composantes fortement ou faiblement connexes du RL-fr. Une étude préliminaire des conséquences d'un tel choix a alors été menée sur une version du RL-fr datant du 12 février 2014.

Le tableau 4.1 fournit quelques informations sur la topologie de la version du RL-fr utilisée. Nous pouvons y voir qu'il était alors partitionnable en 14 212 composantes fortement connexes ou en 4 497 composantes faiblement connexes. Nous avons effectué un regroupement des composantes de chacun de ces types, selon une méthode couplant les deux premières étapes de celle que nous décrivons dans la

[7] Le graphe *E.COLI* ne comporte ni boucle ni arcs multiples.

| | |
|---|---|
| sommets | 21 754 |
| arcs | 40 977 |
| composantes fortement connexes | 14 212 |
| composantes faiblement connexes | 4 497 |

Tab. 4.1 : RL-fr du 12 février 2014

prochaine section. Afin de limiter le temps de cette analyse préliminaire, nous avons focalisé notre attention sur les groupes de composantes ne comportant aucun lien de FL des familles **Syn**, **Anti** et **Contr**, dont nous pensons qu'elles jouent un rôle particulier dans l'organisation du lexique[8]. Pour les composantes fortement connexes, il s'agissait de 15 groupes entre lesquels se répartissaient 71 composantes. Pour les composantes faiblement connexes, il s'agissait de 24 groupes entre lesquels se répartissaient 100 composantes. Dès cette première observation, nous étions invitée à privilégier les composantes faiblement connexes : alors qu'elles étaient initialement moins nombreuses, elles aboutissaient à davantage de regroupements.

Nous avons ensuite observé en détail les groupes, afin de déterminer s'ils contenaient ou non des composantes analogues. Là encore, les groupes de composantes faiblement connexes nous ont paru plus prometteurs. Alors que 22 d'entre eux contenaient un ou plusieurs sous-groupes de composantes analogues, cela était le cas pour seulement 5 groupes de composantes fortement connexes. De plus, quatre des cas de connexions lexicales récurrentes observées étaient représentés, quel que soit le type de composantes sélectionnées.

Une partition en composantes faiblement connexes du RL-fr du 12 février 2014 permettait donc d'accéder à davantage de composantes analogues ne comportant aucun lien de FL des familles **Syn**, **Anti** et **Contr** qu'une partition en composantes fortement connexes, tout en intégrant la quasi-totalité de celles observables dans cette dernière.

### 4.2.5 Données sélectionnées

À la suite de ces considérations, nous avons choisi de réaliser une expérience complète de regroupement de sous-graphes analogues à partir des composantes faiblement connexes de la version du RL-fr du 10 mars 2014.

Le tableau 4.2 présente le pedigree de cette version. Nous pouvons y voir que le RL-fr était alors partitionnable en 4 311 composantes faiblement connexes.

Un peu plus de 60% de ses sommets n'avaient pas encore fait l'objet d'un travail lexicographique et 38% étaient en cours de traitement. Seuls 259 étaient en cours de validation et 4 étaient considérés comme disposant d'une description complète.

Les 42 626 arcs qui le composaient, pour leur part, se répartissent en 36 843 liens

[8]Nous reviendrons sur ce point dans les sections 4.3.4 et 4.4.2.

| | | | |
|---|---|---|---|
| sommets | 21 992 | coefficient d'agrégation | 0,1327 |
| arcs | 42 626 | Distribution des degrés entrants | |
| degré sortant moyen | 1,9383 | a | -2,3977 |
| boucles | 36 | $r^2$ | 0,9397 |
| arcs multiples | 577 | Plus grande composante connexe | |
| arcs symétriques | 19 906 | sommets | 15 302 |
| sommets isolés | 3 226 | arcs | 38 274 |
| composantes faiblement connexes | 4 311 | L | 13,0402 |

TAB. 4.2 : RL-fr du 10 mars 2014

de FL, 1 434 liens de copolysémie et 4 349 liens d'inclusion formelle. Afin de pouvoir prendre en compte ce typage, nous avons associé une étiquette à chaque arc.

- les liens de FL ont été étiquetés à l'aide du préfixe **FL**, suivi de l'identifiant unique de la FL dans la base de données du RL-fr : un lien de nominalisation $\mathbf{S}_0$ devient ainsi **FL21** ;

- les liens de copolysémie ont été étiquetés à l'aide du préfixe **CP**, suivi des identifiants uniques des type et sous-type de copolysémie dans la base de données du RL-fr : un lien de métaphore$_{\text{comme si}}$ devient ainsi **CP1__4** ;

- les liens d'inclusion formelle ont été étiquetés **PH**.

Enfin, concernant les descriptions lexicographiques encapsulées dans les sommets, nous avons observé que 201 d'entre elles ne faisaient mention d'aucune CG. Pas loin de la moitié contenait une étiquette sémantique, mais seulement 44% de ces associations étiquette-lexie disposaient d'un indice de confiance de 100%. Seul un petit tiers contenait une FP. Plus d'un tiers ne disposait d'aucun exemple et 45% en disposaient d'un seul.

Les données disponibles pour cette expérience sont donc plus riches en termes de relations et de descriptions lexicographiques que celles exploitées au chapitre 3. Elles constituent cependant toujours une représentation incomplète du lexique du français.

## 4.3 Regroupement des composantes analogues

Nous allons à présent détailler la méthode de regroupement de microstructures analogues que nous avons mise au point. Après un bref rappel des choix théoriques qui ont guidé son élaboration, nous exposerons chacune des trois étapes qui la constitue. Nous fournirons alors des précisions sur les différents temps d'exécution et sur l'évolution des CFC conservées à l'issue de chacune d'entre elles.

### 4.3.1 Appariement structurel

Rappelons-le, à la suite de Gentner (1983) et Medin et al. (1990), nous considérons le raisonnement analogique comme un appariement structurel. Les lexies s'apparentent alors à des objets disposant d'un certain nombre d'Attributs, disponibles dans leur description lexicographique. Elles entretiennent des Relations, représentées par les arcs du RL-fr.

Dans une telle approche, une analogie s'établit entre une composante source et une composante cible. La « bonne qualité » d'une analogie implique que les Relations présentes dans la composante source soient mises en correspondance avec les Relations de la composante cible. La projection des Attributs est, elle, de moindre importance.

Nous empruntons à Turney (2006) les notions de similarités de Relations et d'Attributs, ainsi que de mesures de celles-ci. Rapportée aux données que nous exploitons, la similarité de Relations entre deux composantes, $C_1$ et $C_2$, dépend du degré de correspondance entre les arcs qui les composent. La mesure de cette similarité est une fonction qui associe les deux composantes à un nombre réel, $sim_r(C_1, C_2) \in \Re$. La similarité d'Attributs, pour sa part, s'établit entre deux lexies $L_1$, $L_2$ et dépend du degré de correspondance entre leurs descriptions lexicographiques. La mesure de cette similarité est une fonction qui associe les deux lexies à un nombre réel, $sim_a(L_1, L_2) \in \Re$.

Tout comme le fait Lepage (2003), nous avons choisi de restreindre l'ensemble des valeurs possibles de $sim_r$ et $sim_a$ en les ramenant à des nombres réels compris entre 0 et 1 ; 0 équivalant à l'absence de similarité, 1 à une similarité complète.

À partir de ces considérations, nous avons choisi d'implémenter le regroupement de microstructures analogues en trois étapes : par isomorphisme, par similarité de Relations et enfin par similarité d'Attributs.

### 4.3.2 Structures isomorphes

La première étape a consisté à regrouper les composantes faiblement connexes, désormais CFC, par structures mathématiques. Chaque CFC a alors été considérée comme un graphe indépendant et comparée aux autres CFC en vue d'établir des ensembles isomorphes. Pour ce faire, nous avons utilisé la librairie python *igraph* et sa fonctionnalité *isomorphic*. Le RL-fr et ses CFC étant orientés, cette fonctionnalité a eu recours à l'algorithme VF2 de Cordella et al. (2001). Cette étape a nécessité un temps d'exécution de 1,5 seconde. Ce temps d'exécution, comme tous ceux mentionnés dans ce chapitre, a été obtenu sur un Mac OS X 10.6.8, muni d'un processeur 3.06 GHz Intel Core 2 Duo et disposant d'une mémoire vive de 4 Go 1067 MHz DDR3..

Deux graphes sont isomorphes s'ils comportent le même nombre de sommets, le même nombre d'arcs et que leurs arcs se répartissent entre les sommets de manière identique. Ainsi, dans la figure 4.9, seuls les deux premiers graphes le sont. Ils correspondent au même motif de taille 3, tels que nous les avons présentés en 4.2.3.

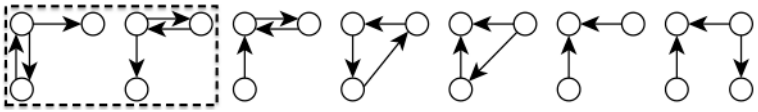


FIG. 4.9 : Exemple d'isomorphisme de graphes

Comme nous l'avons vu précédemment, les sous-graphes comportant moins de trois sommets ne nous intéressent pas. Les 3 226 lexies isolées et les 517 CFC ne contenant que deux sommets ont donc été exclues. Les CFC ne partageant leur structure avec aucune autre, au nombre de 140, ont également été écartées. À l'issue de cette première étape, nous disposions d'un ensemble de 428 CFC, réparties en 36 groupes.

Cet ensemble de CFC rassemblait 1 916 arcs et 1 504 sommets à travers des sous-graphes comportant de deux à six sommets et de deux à seize arcs. Les plus grandes CFC, n'étant pas isomorphes, ont été écartées.

Seulement six de l'ensemble des arcs de ces CFC correspondaient à des liens de copolysémie, quatorze à des liens d'inclusion formelle et la grande majorité, au nombre de 1 896, à des liens de FL. Ces derniers mettaient en œuvre 89 FL distinctes. Le tableau 4.3 présente les cinq plus fréquentes, accompagnées d'exemples. Nous y retrouvons quatre des dix FL les plus fréquentes présentées au chapitre 2. La cinquième, $\mathbf{S_1}$, sert à encoder les relations de dérivation sémantique nominale de premier actant.

| Liens | Étiquette | FL | Exemple |
|---|---|---|---|
| 234 | FL21 | $\mathbf{S_0}$ | FOUILLER → FOUILLE |
| 144 | FL23 | $\mathbf{V_0}$ | FOUILLE → FOUILLER |
| 133 | FL683 | $\mathbf{Syn}_{\subset^{sex}}$ | VOISINE → $\text{VOISIN}_{\text{N}}$ **I** |
| 132 | FL387 | $\mathbf{Syn}_{\supset^{sex}}$ | $\text{VOISIN}_{\text{N}}$ **I** → VOISINE |
| 127 | FL31 | $\mathbf{S_1}$ | FOUILLER → FOUILLEUR |

TAB. 4.3 : FL les plus fréquentes dans les CFC isomorphes

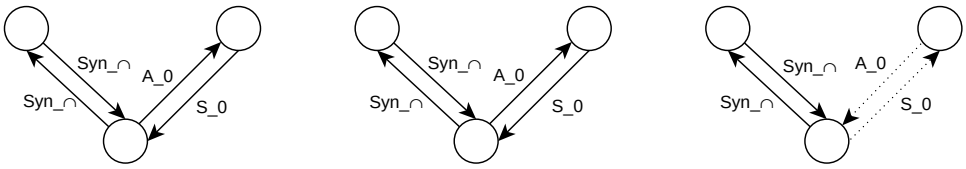
Conformément à notre hypothèse de prédominance de lexies créées à la volée dans les plus petites CFC du RL-fr, 67% des sommets conservés n'avaient pas encore été travaillés et 32% sont en cours de traitement. Seuls deux sommets disposaient d'une description en cours de validation, il s'agissait des lexies $\text{GRAND}_{\text{Adj}}$ **III.1** [*Elle appartient à la grande famille des gadidés.*] et SE DÉPLACER **III.3a** [*Il y a plusieurs façons de se déplacer dans ce livre interactif.*], appartenant à des vocables dont les liens de copolysémie n'avaient alors pas encore été encodés.

### 4.3.3 Similarité de connexions lexicales

La seconde étape a consisté à subdiviser les groupes de CFC isomorphes obtenus précédemment en fonction des Relations en présence. Bien qu'elle représente une étape autonome dans notre raisonnement, elle a été implémentée conjointement à la première étape. Le temps d'exécution de ces deux étapes réunies diminue quelque peu et passe à 1,46 seconde. Cette réduction du temps d'exécution est liée à la diminution du nombre de sous-graphes enregistrés.

Pour réaliser cette subdivision, nous avons comparé l'ensemble des étiquettes d'arcs de chacune des CFC de chaque groupe d'isomorphes. Si les ensembles étaient identiques, nous avons estimé être dans une situation de similarité de Relations complète, équivalent à $sim_r = 1$.

Il s'agit là d'une estimation un peu grossière, car l'orientation des arcs n'est pas prise en compte lors de ce traitement. Aussi, dans des cas tels que celui illustré par la figure 4.10, certaines CFC peuvent être regroupées par similarité de Relations, alors que seulement une partie d'entre elles le sont.




(a) CFC2718 (b) CFC2058 (c) CFC1496

FIG. 4.10 : Exemple de regroupement par similarité de Relations

Dans ce cas précis, la simple observation du typage des arcs et de leur orientation nous permet de savoir que les sommets présents dans ces CFC ne disposent pas tous des mêmes CG fondamentales. Si un lien de synonymie à intersection (**Syn$_\cap$**) impose uniquement que les deux lexies qu'il relie partagent la même partie du discours, les liens de nominalisation (**S$_0$**) sont par définition orientés vers des lexies nominales et ceux d'adjectivation (**A$_0$**) vers des lexies adjectivales. Nous pouvons donc émettre l'hypothèse forte que les CFC 2718 et 2058 sont chacune composées de deux lexies nominales en relation de synonymie à intersection et d'une lexie adjectivale en lien de dérivation syntaxique avec l'une d'entre elles. La CFC 1496, en revanche, doit être composée de deux lexies adjectivales et d'une lexie nominale en lien de dérivation syntaxique avec l'une d'entre elles.

Ces hypothèses sont confirmées par l'observation des sommets. Pour les CFC 2718 et 2058, il s'agit respectivement des lexies TOMBEAU, TOMBE, TOMBAL et PEUPLE**1**, ETHNIE, ETHNIQUE. Pour la CFC 1496, il s'agit des lexies ISRAÉLITE$_{\text{Adj}}$, JUIF$_{\text{Adj}}$**1**, JUDAÏSME. Cette dernière n'est donc pas analogue aux deux autres.

À l'issue de cette étape, 192 CFC ont été conservées, réparties en 47 groupes. Les 236 CFC ne partageant le typage de leurs arcs avec aucune des CFC dont elles

partageaient la structure ont été écartées.

Cet ensemble de CFC rassemblait 769 arcs et 612 sommets à travers des sous-graphes comportant de trois à cinq sommets et de deux à dix arcs. Les CFC isomorphes les plus étendues et mettant en œuvre le plus de Relations ont donc été exclues.

Aucun des groupes obtenus ne comportait de liens de copolysémie. Un seul groupe comportait encore des liens d'inclusion formelle. Il était par ailleurs constitué de CFC composées de trois sommets et de deux arcs, tous deux typés comme étant des liens d'inclusion formelle. Les 46 autres groupes ne mettaient en œuvre plus que 22 FL distinctes. Le tableau 4.4 présente les cinq plus fréquentes. Elles étaient identiques à celles de l'étape précédente, mais leur ordre avait changé. De plus, le couple de lexies que nous avions choisi pour illustrer les FL $\mathbf{Syn}_{\subset^{sex}}$ et $\mathbf{Syn}_{\supset^{sex}}$ n'était plus présent.

| Liens | Étiquette | FL | Exemple |
|---|---|---|---|
| 104 | FL21 | $\mathbf{S_0}$ | FOUILLER → FOUILLE |
| 74 | FL683 | $\mathbf{Syn}_{\subset^{sex}}$ | MIGRAINEUSE → $\text{MIGRAINEUX}_{\text{N}}$ |
| 73 | FL387 | $\mathbf{Syn}_{\supset^{sex}}$ | $\text{MIGRAINEUX}_{\text{N}}$ → MIGRAINEUSE |
| 73 | FL31 | $\mathbf{S_1}$ | FOUILLER → FOUILLEUR |
| 72 | FL23 | $\mathbf{V_0}$ | FOUILLE → FOUILLER |

TAB. 4.4 : FL les plus fréquentes dans les CFC de $sim_r = 1$

Plus aucune lexie en cours de validation n'était présente. La proportion de sommets pour lesquels le travail de description lexicographique n'avait pas débuté était de 68% et ceux pour lesquels le travail était en cours de 32%.

### 4.3.4 Similarité de descriptions lexicographiques

La dernière étape de regroupement analogique a consisté à subdiviser les groupes de CFC en similarité de Relations complète en fonction des lexies qu'elles mettaient en jeu. Le temps d'exécution de cette étape est plus long que celui des deux précédentes. Il dépend du nombre de groupes, du nombre de sous-graphes qu'ils contiennent et du nombre de sommets qui les composent. Dans le cas de notre expérience, près de la moitié des groupes ne contenaient que deux CFC et six en rassemblaient entre dix et quinze, qui étaient toutes composées de trois sommets. Leur subdivision a été réalisée en 2 minutes et 34 secondes.

Comme nous l'avons souligné, l'orientation des Relations n'avait pas été prise en compte dans le précédent regroupement. Nous pensions qu'une ultime étape, de regroupement par similarité d'Attributs des lexies, permettrait de remédier à ce manque. Ainsi, en comparant l'ensemble des lexies de la CFC 2718 à celles de la CFC 2058, trois couples de lexies en situation de similarité d'Attributs complète seraient obtenus : (TOMBEAU, PEUPLE**I**), (TOMBE, ETHNIE) et (TOMBAL, ETHNIQUE). En re-

vanche, en comparant l'ensemble des lexies de la CFC 2718 à celle de la CFC 1496, seuls deux couples de lexies le seraient : (TOMBE, JUDAISME) et (TOMBAL,$\text{JUIF}_{\text{Adj}}$ **1**). Les CFC 2718 et 2058 seraient alors regroupées, en tant que CFC analogues relevant d'une même configuration de dérivations lexicales, tandis que la CFC 1496 se retrouverait isolée.

Nous avons présenté, dans le chapitre 2, l'ensemble des éléments de description lexicographique encapsulés dans les sommets du RL-fr comme un ensemble d'Attributs disponibles pour la comparaison des lexies. Certains — l'appartenance à un vocable et la flexion morphologique — ont d'ores et déjà été écartés. Les autres fournissent des informations de natures diverses et ne nous semblent pas tous pertinents pour regrouper les CFC relevant d'un même modèle de dérivation.

Les étiquettes sémantiques permettent d'établir des rapports entre lexies selon l'organisation du lexique en espaces sémantiques. Nous ignorons pour l'heure si les cas de dérivations lexicales qui nous intéressent dépendent de cette organisation et si une même configuration peut ou non être observée dans des espaces distincts. Afin de ne pas contraindre nos observations selon cette dimension, nous avons donc choisi de ne pas prendre en compte cet élément de description.

D'une façon comparable, l'ensemble des phrasèmes formés à partir d'une lexie fournit des informations sur l'organisation phraséologique du lexique. Il s'agit là encore d'une dimension selon laquelle nous ne souhaitons pas établir la similarité entre sous-graphes. Cet élément de description n'a donc pas non plus été retenu.

Les FP, en fournissant le nombre d'actants de chaque lexie, nous semblent davantage pertinentes. Les données disponibles n'ont cependant pas permis leur exploitation lors cette expérience. En effet, seuls 28 des 612 sommets lexicaux conservés à cette étape disposaient de cet élément de description.

Les mêmes difficultés ont été rencontrées pour les exemples lexicographiques, accentuées par une forte disparité. Tandis que l'un des sommets lexicaux conservés était associé à 16 exemples distincts, 514 d'entre eux n'en disposaient d'aucun[9].

Deux types d'éléments de description restaient alors à notre disposition pour chaque lexie : ses caractéristiques grammaticales et les liens de FL dont elle était source ou cible.

Les caractéristiques grammaticales encodées dans le RL-fr sont variées. Il s'agit de caractéristiques fondamentales[10], de marques d'usage langagier, stylistique et rhétorique, de caractéristiques formelles, de positions syntaxiques et d'informations de linéarisation. Chacune de ces informations ne semble pas pertinente pour déterminer si les connexions lexicales entre lexies sont analogues. Par exemple, la différence

[9]Rappellons qu'un certain nombre d'exemples lexicographiques ont été enregistrés dans la base de données du RL-fr avant l'implémentation des fonctionnalités qui leur sont consacrées. Ils sont disponibles sous forme de commentaires et sont exclus de nos statistiques.

[10]Les caractéristiques fondamentales d'une lexie sont principalement sa partie du discours et son genre, s'il s'agit d'une lexie nominale.

entre caractéristiques fondamentales de genre pour deux lexies nominales n'est significative que dans des cas particuliers, comme la dérivation entre un nom de métier masculin et son équivalent féminin. Nous avons donc souhaité concentrer notre attention sur les parties du discours.

La granularité de ces dernières — 50 parties du discours de surface et 9 parties du discours profondes [11]— risquait cependant de nous amener à considérer comme différentes des CFC que nous aurions souhaité conserver dans un même groupe. Nous avons donc eu recours à un artifice élaboré en collaboration avec les lexicographes : un ensemble de 13 méta-parties du discours, désormais méta-pdd, auxquelles a été rapportée chacune des 59 parties du discours existantes — `Nom`, `Adjectif`, `Verbe`, `Adverbe`, `Clausatif`, `Numéral`, `Déterminant`, `Pronom`, `Préposition`, `Conjonction`, `Affixe`, `Adjectif/Adverbe`, `Adverbe/Adjectif`.

Les liens de FL associés à une lexie, pour leur part, nous ont semblé être des éléments essentiels pour effectuer les regroupements souhaités. En effet, ils fournissent des informations sur la combinatoire lexicale et le paradigme de chaque lexie. Ils permettent ainsi de se faire une idée sur son rôle dans l'organisation du lexique. Une lexie verbale comme FAIRE **II.1** [*Il fait du ping-pong.*], par exemple, joue un rôle carrefour. Dans la version du RL-fr du 10 mars 2014, elle n'était associée à aucun lien de FL sortant, mais a 106 liens de FL entrants — dont 82 rendant compte de son utilisation en tant que verbe support et 19 en tant que verbe de réalisation. Elle se distinguait en cela de lexies verbales plus classiques, encodant des liens de FL sortants et bien moins de liens entrants. La description de KIDNAPPER **I** [*Les ravisseurs ont kidnappé son fils, et demandent une rançon de 100 000 euros.*] comportait, par exemple, 14 liens de FL sortants — encodant des relations de synonymie, de nominalisation et de dérivation sémantique nominale actancielle — et quatre liens de FL entrants — encodant des relations de synonymie, de verbalisation et de causation.

La comparaison de telles lexies nous a conduit à penser que trois propriétés relevant de cet élément de description méritaient d'être considérées comme pertinentes : la nature des liens de FL sortants, la nature des liens de FL entrants et le rapport arithmétique entre le nombre de liens entrants et le nombre de liens sortants. Cependant, la granularité des FL risquait, ici aussi, de nous amener à différencier des CFC qui ne devraient pas l'être. En effet, nous pouvions alors en dénombrer 673 distinctes dans l'ensemble du RL-fr. Nous avons donc choisi d'établir la comparaison des natures de liens de FL au niveau de leurs familles.

Nous avons également choisi d'exclure les familles permettant de rendre compte de relations de synonymie (**Syn**), d'antonymie (**Anti**) et de contrastivité (**Contr**). Nous avons estimé que les liens relevant de ces familles amèneraient à des distinctions inappropriées. Ainsi, deux CFC en similarité de Relations complète mettant en jeu l'une la lexie FRÉQUEMMENT et l'autre la lexie EXTRÊMEMENT auraient été considérées comme différentes, car la lexie FRÉQUEMMENT entretenait alors une re-

[11] Les parties du discours profondes se distinguent des parties du discours de surface. Ainsi, un nom commun (partie du discours de surface) comme la lexie BŒUF **IV** [*Ryan Gosling lui fait un effet bœuf.*] qui a un emploi appositif, possède la valence passive d'un adjectif (partie du discours profonde). Pour une introduction détaillée de ces notions, voir Mel'čuk (2006).

lation de synonymie avec SOUVENT, tandis que la lexie EXTRÊMEMENT ne comptait aucun synonyme.

Nous avons finalement associé à chaque lexie un ensemble d'Attributs constitué de la manière suivante :

- un Attribut rendant compte de sa méta-pdd ;
- autant d'Attributs FLout que de familles de FL en jeu dans l'ensemble de ses liens de FL sortants ;
- autant d'Attributs FLin que de familles de FL en jeu dans l'ensemble de ses liens de FL entrants ;
- un Attribut rendant compte du rapport arithmétique entre le nombre de liens entrants et le nombre de liens sortants, valant **out**+ en cas de supériorité numérique des liens sortants, **in**+ en cas de supériorité numérique des liens entrants ou **in**=**out** en cas d'égalité.

L'implémentation de cette association est identique à celle présentée dans le chapitre 3. Nous avons généré un fichier RDF N3 par lexie à l'aide d'un script PHP interrogeant la base de données du RL-fr. La figure 4.11 montre un exemple d'un tel fichier, dans le cas de la lexie FOUILLER. Le nom de la lexie y est fourni sous forme de commentaire. Tout comme les deux de préfixes, la ligne qui lui est consacrée n'est prise en compte dans aucun traitement.

```
@prefix rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>.
@prefix rlf:<http://relief.atilf.fr/RLF#>.
# rlf:lexie rlf:name rlf:-fouiller---.
rlf:lexie a rlf:V.
rlf:FL11 a rlf:FLout.
rlf:FL15 a rlf:FLout.
rlf:FL12 a rlf:FLin.
rlf:lexie a rlf:+out.
```

FIG. 4.11 : Fichier RDF N3 de la lexie FOUILLER

Le calcul de similarité d'Attributs entre lexies est pour sa part plus simple que celui présenté au chapitre 3. Chaque Attribut est ici unique et son ordre d'apparition dans le fichier RDF N3 n'est pas important. L'implémentation d'une distance d'édition a donc été abandonnée. Elle a été remplacée par une simple comparaison d'ensemble, de la manière suivante :

$$sim_a(Lexie_1, Lexie_2) = \frac{2 \times nbr\ d'attributs\ communs}{nbr\ d'attributs\ Lexie_1 + nbr\ d'attributs\ Lexie_2}$$

Nous avons ainsi mesuré la similarité des lexies en jeu dans chacun des 47 groupes de CFC en situation de similarité de Relations complète. Nous nous sommes ensuite appuyée sur les résultats de cette mesure pour effectuer une dernière subdivision et sélectionner les cas de CFC analogues.

Prenons le cas d'un groupe composé des deux CFC de trois lexies, $CFC_1$ et $CFC_2$. Chacun des ensembles d'Attributs des lexies de $CFC_1$ a être comparé à chacun des ensembles d'Attributs des lexies de $CFC_2$. Il s'agissait donc d'effectuer neuf mesures de $sim_a$. Si exactement trois de ces mesures aboutissaient au résultat $sim_a = 1$, les CFC ont été considérées comme étant analogues et maintenues dans un seul groupe. Si plus de trois mesures aboutissaient au résultat $sim_a = 1$, le groupe a été maintenu, mais la question de l'analogie des CFC est restée en suspens. Enfin, si moins de trois mesures aboutissaient au résultat $sim_a = 1$, les CFC ont été considérées comme non analogues et chacune s'est retrouvée isolée.

Dans le cas de groupes composés de plus de deux CFC, plusieurs regroupements pouvaient être envisagés. En effet, une même CFC peut partager un nombre différent de $sim_a = 1$ avec chacune des CFC de son groupe initial. Nous avons alors décidé de regrouper les CFC par nombre maximal de $sim_a = 1$. Une fois les CFC partageant le plus de $sim_a = 1$ regroupées, le nombre de $sim_a = 1$ partagées par les CFC restantes a considéré, etc.

À l'issue de cette dernière étape, nous disposions de 92 CFC réparties en 24 groupes de CFC analogues. Les autres CFC se répartissaient en 23 groupes de 86 CFC pour lesquels la question de l'analogie n'était pas tranchée et 14 CFC isolées.

### 4.3.5 Ensembles de proportions analogiques

La méthode que nous proposons ici se différencie de celle que nous avons exploitée dans l'expérience présentée dans le chapitre 3 sur plusieurs points :

Plutôt que de nous appuyer sur une mesure de similarité d'Attributs entre descriptions de lexies prenant en compte tous les éléments disponibles, nous avons choisi de sélectionner ceux qui nous semblaient les plus pertinents. Nous nous dispensons ainsi de la nécessité d'établir un seuil. Les Attributs comparés constituent le socle minimal commun pour que deux lexies occupent la même place dans une configuration de dérivations lexicales. Ils nous permettent, en quelque sorte, de contraindre davantage la nature des objets entre lesquels nous cherchons à établir une conformité de rapports.

Alors que nous avions précédemment eu recours au score de Turney pour pallier l'absence de mesure de similarité de Relations, nous exploitons à présent la structure mathématique des sous-graphes et le typage de leurs liens à cette fin.

Rappelons-le, nous pensons qu'une analogie entre deux sous-graphes se base sur un ensemble de proportions analogiques entre leurs sommets. Ainsi, en énonçant que les deux CFC composées de trois sommets de la figure 4.12 sont analogues, nous sous-entendons l'existence de trois classes de huit analogies équivalentes.

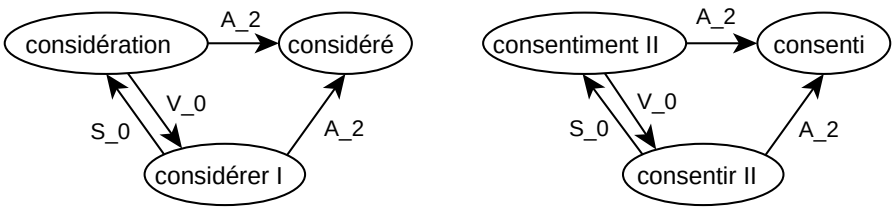



FIG. 4.12 : Exemple de CFC analogues

Chacune de ces classes correspond à l'une des proportions suivante :

CONSIDÉRATION : CONSIDÉRER I :: CONSENTEMENT II : CONSENTIR II,
CONSIDÉRÉ : CONSIDÉRATION I :: CONSENTI II : CONSENTEMENT II,
CONSIDÉRER I : CONSIDÉRÉ :: CONSENTIR II : CONSENTI.

Si nous nous intéressons à la première d'entre elles, nous pouvons en effet constater que, bien que notre mesure de $sim_r$ ait été établie sur les sous-graphes dans leur globalité, les Relations encodées par les lexicographes entre les lexies CONSIDÉRATION et CONSIDÉRER I sont identiques à celles encodées entre les lexies CONSENTEMENT II et CONSENTIR II. De plus, si ces CFC sont considérées comme analogues, c'est que nous avons établi que les similarités d'Attributs étaient complètes, aussi bien entre CONSIDÉRATION et CONSENTEMENT II, qu'entre CONSIDÉRER I et CONSENTIR II. Dès lors, l'écart de similarité d'Attributs entre CONSIDÉRATION et CONSIDÉRER I est nécessairement identique à celui existant entre CONSENTEMENT II et CONSENTIR II.

Le problème de l'énonciation des rapports en jeu dans une partie des analogies équivalentes entre lexies — que nous avons évoqué au chapitre 2, section 2.2.2.1 — n'est cependant toujours pas réglé. Rappelons-le, selon Lepage (2003), si « A est à B ce que C est à D », alors « A est à C ce que B est à D ». Or, à ce stade, nous ne sommes toujours pas en mesure d'énoncer de façon satisfaisante ce que « CONSIDÉRATION est à CONSENTEMENT II », pas plus que ce que « CONSIDÉRER I est à CONSENTIR II ». Si nous avons pris le parti de considérer qu'il s'agit d'objets de même nature qui occupent une place identique dans une configuration de dérivations lexicales, l'énonciation de ces rapports selon une dimension plus large, relevant de l'organisation générale du lexique, ne nous est pas encore accessible.

## 4.4 Analyse des résultats

Afin de vérifier la pertinence de notre méthodologie et de nous faire une idée sur l'exploitation possible des résultats obtenus, nous avons observé en détail l'ensemble des CFC qui se trouvaient regroupées et isolées à l'issue de la dernière étape.

### 4.4.1 Groupes de composantes analogues

Notre premier constat a été que chacun des 24 groupes de CFC analogues que nous avions obtenus étaient correctement constitués. Ils comportaient de deux à six

CFC analogues, mettant toutes en jeu des ensembles de trois lexies, reliées exclusivement par des liens de FL. Plus de la moitié des groupes ne comportaient que deux CFC. Les autres en comportaient entre trois et quatorze.

#### 4.4.1.1 Topologie des composantes analogues

Parmi les 276 sommets qui composaient ces CFC, 94 étaient en cours de traitement lexicographique et 182 avaient juste été créés. Les arcs, au nombre de 301, ne mettaient plus en œuvre que 16 FL distinctes. Le tableau 4.5 présente les cinq plus fréquentes. Seules les deux premières sont identiques à celles de l'étape précédente. Les trois nouvelles servent à encoder des relations d'adjectivation ($\mathbf{A_0}$), d'adverbalisation ($\mathbf{Adv_0}$) et de dérivation sémantique adjectivale de premier actant ($\mathbf{A_1}$). Les liens encodant des relations de dérivations syntaxiques sont majoritaires. À elles seules, les quatre présentes ici couvrent plus de 63% des arcs.

| Liens | Étiquette | FL | Exemple |
|---|---|---|---|
| 72 | FL21 | $\mathbf{S_0}$ | FOUILLER → FOUILLE |
| 48 | FL23 | $\mathbf{V_0}$ | FOUILLE → FOUILLER |
| 37 | FL28 | $\mathbf{A_0}$ | ÉTRANGEMENT → ÉTRANGE |
| 34 | FL29 | $\mathbf{Adv_0}$ | ÉTRANGE → ÉTRANGEMENT |
| 22 | FL104 | $\mathbf{A_1}$ | ÉTRANGETÉ → ÉTRANGE |

TAB. 4.5 : FL les plus fréquentes dans les CFC analogues

Cinq des groupes constitués avaient la particularité de rassembler des CFC comportant deux lexies de même méta-pdd. La comparaison des Attributs des lexies de ces CFC deux à deux aboutissait cependant dans tous les cas à seulement trois $sim_a = 1$. Ces similarités complètes correspondaient aux couples de lexies occupant une place identique dans la structure des CFC.

Ainsi, les CFC de la figure 4.13 présentent la même configuration de dérivations lexicales. Elles mettent en jeu des relations de verbalisation ($\mathbf{V_0}$), de nominalisation ($\mathbf{S_0}$) et de dérivation sémantique nominale de premier actant ($\mathbf{S_1}$). Elles comportent chacune un verbe et deux noms. Dans ce cas précis, trois similarités d'Attributs complètes ont été comptabilisées, entre les lexies des couples (SUCCÉDER **I**, FOUILLER), (SUCCESSION **I**, FOUILLE) et (SUCCESSEUR, FOUILLEUR).

Quelles que soient les CFC regroupées, les lexies qu'elles contiennent sont peu décrites dans le RL-fr. Il n'a donc pas été possible de les exploiter pour définir des profils de lexies susceptibles de déclencher des relations particulières.

En revanche, certains groupes de CFC présentaient des structures imbricables. Cette particularité nous a amenée à nous interroger sur la granularité des configurations de dérivations lexicales. Faut-il chercher à établir les modèles les plus denses possibles et exploiter les imbrications de CFC analogues pour enrichir automatiquement les CFC comportant le moins de liens de FL ?

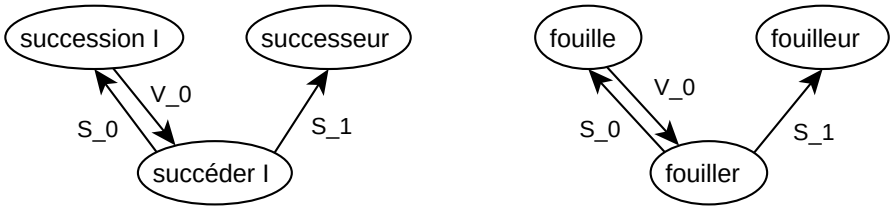
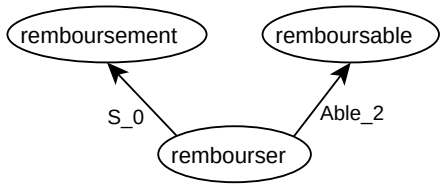
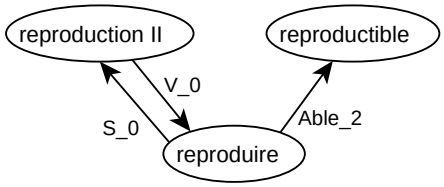
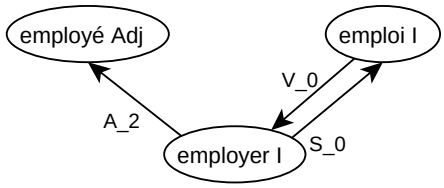
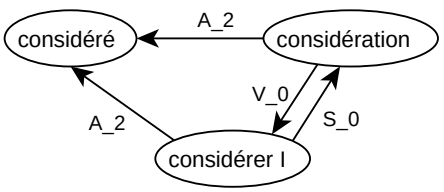



FIG. 4.13 : Exemple de CFC analogues avec deux lexies de même méta-pdd

#### 4.4.1.2 Imbrication de groupes de composantes analogues

| groupe 1 | groupe 2 | ajouts |
|---|---|---|
|  |  | $\mathbf{V_0}$ |
|  |  | $\mathbf{A_2}$ |



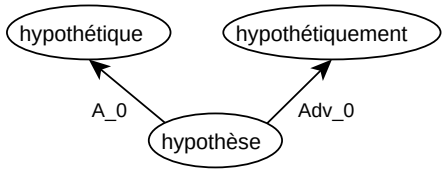
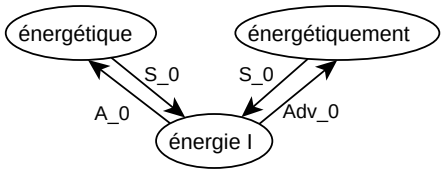
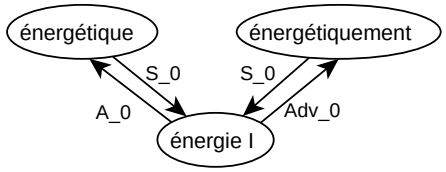
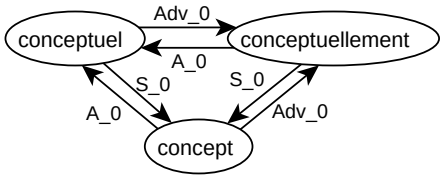
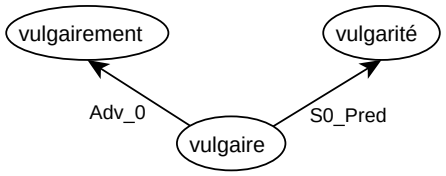
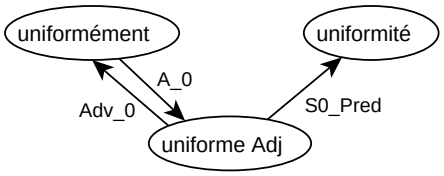
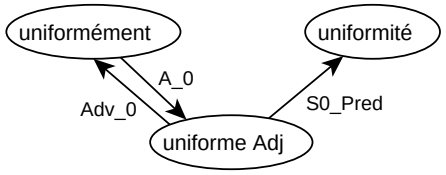
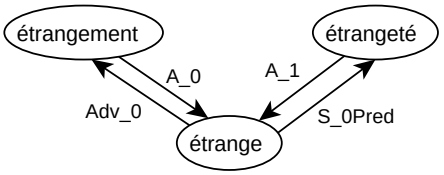
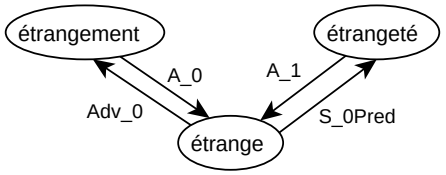
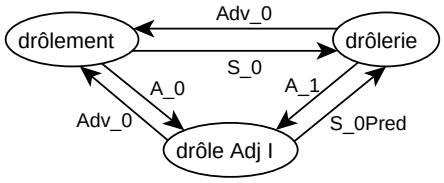
TAB. 4.6 : Imbrications avec méta-pdd `Adjectif`, `Nom` et `Verbe`

Le tableau 4.6 présente un premier ensemble d'imbrications observées et les suggestions d'ajout de liens qui en découlent. Il concerne des CFC dont les lexies ont pour méta-pdd `Adjectif`, `Nom` et `Verbe`.

Nous pouvons y observer deux imbrications distinctes. Elles mettent toutes deux en jeu des relations de nominalisation ($\mathbf{S_0}$) et de verbalisation ($\mathbf{V_0}$). Dans le cas de la première, une relation de dérivation sémantique adjectivale de deuxième actant ($\mathbf{A_2}$) s'y ajoute. Tandis que le premier groupe impliqué dans cette imbrication ne comporte que deux CFC, le second en comporte neuf. De plus, comme nous l'avons déjà évoqué, la cooccurrence d'une FL de nominalisation $\mathbf{S_0}$ et d'une FL de verbalisation $\mathbf{V_0}$ est un cas connu de FL inverses. L'enrichissement des CFC du premier groupe à partir de celles du second est donc souhaitable.

Dans la seconde imbrication, le lien supplémentaire rend compte d'une relation de dérivation sémantique adjectivale potentielle de deuxième actant ($\mathbf{S_2}$). Les deux groupes qui y sont impliqués comportent chacun deux CFC. Nous ne disposons donc pas d'un écart de fréquence permettant de justifier l'enrichissement automatique des

CFC du premier groupe à partir de celles du second. Le recours à l'expertise des lexicographes est nécessaire.

| groupe 1 | groupe 2 | ajouts |
|---|---|---|
|  |  | 2 **$S_0$** |
|  |  | **$A_0$**<br>**$Adv_0$** |
|  |  | **$A_0$** |
|  |  | **$A_1$** |
|  |  | **$S_0$**<br>**$Adv_0$** |

TAB. 4.7 : Imbrications avec méta-pdd `Adjectif`, `Nom` et `Adverbe`

Le tableau 4.7 présente un second ensemble d'imbrications observées. Il concerne des CFC dont les lexies ont pour méta-pdd `Adjectif`, `Nom` et `Adverbe`. Les imbrications y sont plus nombreuses. Elles s'organisent en cascade et nous pouvons considérer qu'elles se divisent en deux chaînes distinctes.

La première de ces chaînes met en jeu des relations de nominalisation (**$S_0$**), d'adjectivation (**$A_0$**) et d'adverbialisation (**$Adv_0$**). Elle débute par une imbrication qui met en concurrence un groupe de deux CFC ne comportant que deux arcs et un groupe de cinq CFC comportant quatre arcs. Les ajouts de liens suggérés par cette imbrication rétablissent des cas connus de FL inverses. L'imbrication suivante implique un troisième groupe. Celui-ci n'est composé que de trois CFC. Cependant, il permet d'envisager un enrichissement qui suit la même logique que celle des FL inverses.

La seconde chaîne met en jeu des relations d'adverbialisation (**$Adv_0$**), de nominalisation simple (**$S_0$**) et prédicative (**$S_0Pred$**), d'adjectivation (**$A_0$**) et de dérivation sémantique adjectivale de premier actant (**$A_1$**). Elle implique des groupes qui comportent, par ordre d'apparition, trois, sept, sept et deux CFC. Le premier lien dont elle suggère l'ajout rétablit un cas connu de FL inverses et les deux derniers suivent une logique comparable à celle-ci. Au contraire, le lien intermédiaire, bien que présent dans neuf CFC et semblant suivre cette même logique, est connu pour être un cas problématique.

Nous avons consulté l'équipe de lexicographes pour savoir si les CFC les moins denses étaient toujours valides une fois enrichies. Un seul des cas que nous leur avons présentés a été rejeté. Il s'agit du résultat de la seconde chaîne d'imbrications concernant des groupes de lexies ayant pour méta-ppd `Adjectif`, `Nom` et `Adverbe`. L'ajout d'une relation de dérivation sémantique adjectivale de premier actant (**$A_1$**) est considéré comme une erreur. Cette observation nous met en garde contre la propagation automatique de mauvais liens. À ce stade, aucune corrélation n'a être établie entre l'écart de fréquence des groupes impliqués et la validité d'une imbrication. Nous n'écartons cependant pas la possibilité que l'observation d'un plus grand nombre de sous-graphes le permette.

#### 4.4.1.3 Configurations de dérivations syntaxiques

Nous avons souligné la forte concentration de liens de dérivation syntaxique dans les groupes de CFC analogues. De tels liens ont principalement été encodés au cours de la phase de constitution de la nomenclature directement induite, que nous avons présenté au chapitre 2, section 2.1.3. Bien que la dérivation dont il est ici question soit exclusivement syntaxique, nous avons observé au fil de notre analyse que les lexies regroupées dans une même CFC étaient très fréquemment morphologiquement apparentées.

Nous avons présenté, dans le chapitre 1, les travaux de Hathout (2009, 2011) et la ressource Morphonette associée. Constituée par informatisation de l'analogie formelle et sémantique à partir du TLFi, cette ressource regroupe 29 310 mots[12], 96 107 relations de type famille morphologique dérivationnelle et 1 160 098 relations de type série dérivationnelle. Rappelons-le, une famille morphologique regroupe un ensemble de mots qui partagent des propriétés formelles et sémantiques les plus spécifiques possibles, tels que « *modifiable, modificateur, modifier, modification* ». Une série morphologique, pour sa part, regroupe un ensemble de mots qui partagent des

[12]Les mots dont il est question ici sont des associations de forme et de partie du discours.

propriétés sémantiques et formelles générales et participent au plus grand nombre possible d'analogies formelles qui impliquent les autres membres de la série, tels que « *modifiable, fructifiable, rectifiable, sanctifiable* ».

Il nous a semblé intéressant de confronter les groupes de CFC analogues mettant exclusivement en œuvre des relations de dérivations syntaxiques avec la ressource Morphonette[13]. Le tableau 4.8 présente ces groupes. Ils sont au nombre de six, comportent de deux à sept CFC, chacune composée de trois sommets et de deux à six arcs. Ils rassemblent au total 22 CFC, 66 sommets et 77 arcs.

| nombre de CFC | nombre d'arcs | FL en jeu |
|---|---|---|
| 3 | 2 | $\mathbf{Adv_0}$, $\mathbf{S_0Pred}$ |
| 2 | 2 | $\mathbf{A_0}$, $\mathbf{Adv_0}$ |
| 7 | 3 | $\mathbf{A_0}$, $\mathbf{Adv_0}$, $\mathbf{S_0Pred}$ |
| 5 | 4 | $\mathbf{S_0}$, $\mathbf{S_0}$, $\mathbf{A_0}$, $\mathbf{Adv_0}$ |
| 2 | 4 | $\mathbf{S_0}$, $\mathbf{V_0}$, $\mathbf{V_0}$, $\mathbf{A_0}$ |
| 3 | 6 | $\mathbf{S_0}$, $\mathbf{S_0}$, $\mathbf{A_0}$, $\mathbf{A_0}$, $\mathbf{Adv_0}$, $\mathbf{Adv_0}$ |

TAB. 4.8 : Groupes de structures de dérivations syntaxiques analogues

Cette confrontation a été effectuée au regard de deux interrogations. D'une part, nous avons cherché à déterminer si les sommets de chacune des CFC étaient répertoriés dans Morphonette comme membres d'une même famille. D'autre part, nous avons regardé si les lexies qui occupaient une place identique dans les groupes de CFC analogues étaient répertoriées comme membres d'une même série.

Parmi les 22 CFC, nous n'avons observé que quatre cas pour lesquels toutes les lexies étaient présentes dans Morphonette. Deux d'entre eux concernaient des lexies répertoriées comme appartenant à la même famille morphologique : (GÉOGRAPHIE, GÉOGRAPHIQUE, GÉOGRAPHIQUEMENT) et (AUTOMATIQUE$_{\text{Adj}}$, AUTOMATISME, AUTOMATIQUEMENT). Un troisième cas associait les lexies ÉNUMÉRATION et ÉNUMÉRATIF, répertoriées comme appartenant à la même famille, à la lexie ÉNUMÉRER, présente uniquement en tant que membre de séries. Le dernier cas comportait les lexies ÉNERGIE**I** et ÉNERGÉTIQUE, répertoriées dans des familles distinctes et la lexie ÉNERGÉTIQUEMENT, uniquement membre de séries.

Nous avons comptabilisé cinq CFC dont les lexies étaient toutes absentes de Morphonette. Elles forment les ensembles de lexies (SPLENDIDE, SPLENDEUR, SPLENDIDEMENT), (NÉCESSAIRE$_{\text{Adj}}$ **1**, NÉCESSITÉ, NÉCESSAIREMENT), (HYPOTHÈSE, HYPOTHÉTIQUE, HYPOTHÉTIQUEMENT), (BIZARRE, BIZARRERIE, BIZARREMENT) et (BRUSQUE, BRUSQUERIE, BRUSQUEMENT) qui nous paraissent répondre aux critères d'appartenance à une même famille morphologique.

[13]Nous avons utilisé la version 0.1 de Morphonette, téléchargée à l'adresse `http ://redac.univ-tlse2.fr/lexiques/morphonette.html`.

Dix CFC étaient composées de deux lexies présentes dans Morphonette et d'une absente. Parmi celles-ci, nous avons observé huit couples de lexies présentes et répertoriées comme appartenant à une même famille. Les lexies associées à ces couples étaient des noms, au nombre de cinq, et des adjectifs, au nombre de trois. Par exemple, la lexie CONCEPT était associée au couple (CONCEPTUEL, CONCEPTUELLEMENT) et la lexie SUBTIL au couple (SUBTILITÉ, SUBTILEMENT). Toutes paraissaient être de bonnes candidates pour compléter les familles existantes. Les deux CFC restantes étaient composée d'au moins une lexie qui n'était répertoriée dans Morphonette qu'en tant que membre de séries morphologiques.

Enfin, trois CFC ne comportaient qu'une seule lexie répertoriée dans Morphonette. Elles regroupaient les lexies (CÉRÉBRAL, CÉRÉBRALEMENT, CERVEAU **I**), (SACRIFIER, SACRIFICE, SACRIFICIEL) et (BASSEMENT, $\text{BAS}_{\text{Adj}}$ **III**, BASSESSE).

L'examen des lexies occupant une place identique n'a, pour sa part, donné aucun résultat. Aucun cas de séries morphologiques dérivationnelles n'a pu être comptabilisé dans nos groupes de CFC analogues.

Ces observations nous amènent à considérer que l'approche de la dérivation morphologique par informatisation de l'analogie formelle et sémantique de Hathout (2009, 2011) et notre approche de la dérivation syntaxique exploitant l'encodage explicite de relations par les lexicographes sont complémentaires. Malgré la faible quantité de données analysées, nous avons vu que nous étions en mesure de proposer des candidats pour compléter les familles morphologiques répertoriées dans Morphonette. Il serait également intéressant de considérer les choses du point de vue inverse. Nous pourrions alors chercher à établir si les informations contenues dans Morphonette peuvent être mises à disposition des lexicographes sous une forme adaptée et si une telle fonctionnalité leur serait profitable.

### 4.4.2 Groupes de composantes à l'analogie incertaine

Après avoir analysé les groupes de CFC analogues, nous nous sommes intéressée aux 23 groupes de CFC dont la question de l'analogie était restée en suspens. Ils étaient composés de CFC mettant en jeu des ensembles de trois à cinq lexies, reliées par deux à dix arcs. Ici aussi, plus de la moitié des groupes ne comportaient que deux CFC. Les autres en comportaient entre trois et quinze.

Dix des 422 arcs répartis dans ces CFC rendaient compte de relations d'inclusion formelle. Les autres mettaient en œuvre 16 FL distinctes. Le tableau 4.9 présente les cinq plus fréquentes.

Nous pouvons y observer qu'un peu plus d'un tiers des arcs encodaient des relations de quasi-synonymie relative au sexe ($\mathbf{Syn}_{\subset^{sex}}$, $\mathbf{Syn}_{\supset^{sex}}$) et qu'un tiers encodaient des dérivations sémantiques nominales actancielles du premier actant ($\mathbf{S_1}$, $\mathbf{S_1Pred}$, $\mathbf{S_1}^{\mathbf{usual}}$).

Les sommets entre lesquels étaient établies ces relations étaient au nombre de 294. Seuls 91 d'entre eux correspondaient à des lexies en cours de traitement lexicographique, les 203 autres avaient tout juste été créés.

| Liens | Étiquette | FL | Exemple |
|---|---|---|---|
| 71 | FL683 | $\mathbf{Syn}_{\subset^{sex}}$ | AMBITIEUX$_{\text{N}}$ → AMBITIEUSE |
| 71 | FL387 | $\mathbf{Syn}_{\supset^{sex}}$ | AMBITIEUSE → AMBITIEUX$_{\text{N}}$ |
| 66 | FL31 | $\mathbf{S_1}$ | CAISSE **II** → CAISSIÈRE |
| 40 | FL594 | $\mathbf{S_1Pred}$ | ANGLAIS$_{\text{Adj}}$ **2** → ANGLAISE |
| 32 | FL30 | $\mathbf{S_1}^{\mathbf{usual}}$ | FORMER **III** → FORMATEUR$_{\text{N}}$ |

TAB. 4.9 : FL les plus fréquentes dans les CFC à l'analogie incertaine

Nous avons commencé par nous intéresser aux groupes de CFC dont les arcs encodaient des liens d'inclusion formelle. Ils étaient au nombre de deux, l'un comportant trois CFC, l'autre deux. Bien qu'ils sortent du cadre de la dérivation lexicale qui nous intéresse, nous avons remarqué que chacun de ces groupes correspondait à une structure syntaxique de locution nominale particulière : **N + de + N** pour l'un, **Adj + N** pour l'autre. La figure 4.14 illustre ces structures.

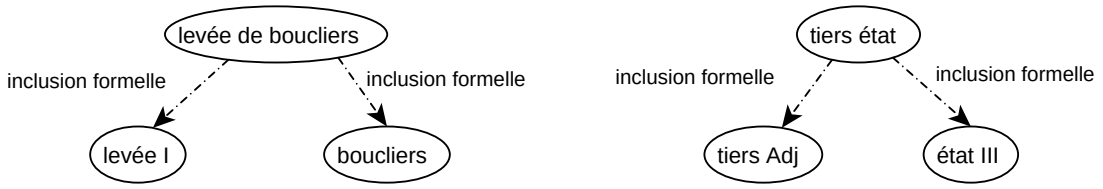



FIG. 4.14 : Exemples de CFC comportant des relations d'inclusion formelle

Parmi les autres groupes, seuls trois se sont avérés contenir des CFC non analogues. Il s'agissait de groupes de deux CFC, composées dans deux cas de quatre sommets et cinq arcs et dans un cas de trois sommets et quatre arcs. Ils mettaient tous en jeu des relations de quasi-synonymie relative au sexe. Cinq $sim_a = 1$ ont été comptabilisées entre les CFC de quatre sommets et quatre entre celles de trois sommets.

Les CFC du premier groupe mettaient en jeu les lexies (DISTRACTION **II**, DISTRAIT$_{\text{N}}$, DISTRAITE, DISTRAIT$_{\text{Adj}}$) et (SE RÉSIGNER, RÉSIGNÉ$_{\text{N}}$, RÉSIGNÉE, RÉSIGNÉ$_{\text{Adj}}$). Nous avons immédiatement constaté que ces lexies ne partageaient pas toutes les mêmes méta-pdd deux à deux. Elles ne relevaient donc pas d'une même configuration de dérivation lexicale.

Les CFC du second groupe mettaient en jeu les lexies (TÉMOIGNER, TÉMOIGNAGE **I.2**, TÉMOIN$_{\text{N,fém}}$, TÉMOIN$_{\text{N,masc}}$) et (DONNER **I.3**, DON **I.2**, DONATEUR, DONATRICE). Ces lexies auraient dû partager les mêmes méta-pdd, mais la lexie DONNER **I.3** avait accidentellement été associée aux CG `nom commun` et `masc.` Sans cette erreur, corrigée depuis, six $sim_a = 1$ auraient été comptabilisées pour ce groupe.

Les CFC du troisième groupe mettaient en jeu les lexies (BLOND **II.a**, BLONDE **I**, BLOND **II.b**) et (PARACHUTE, $\text{PARACHUTISTE}_{\text{N,fém}}$, $\text{PARACHUTISTE}_{\text{N,masc}}$). A priori, ces lexies semblaient partager les mêmes méta-pdd. Cependant, le vocable BLOND regroupe à la fois des noms et des adjectifs[14]. Les acceptions en présence relevaient chacune d'une de ces méta-pdd.

L'ensemble des 18 groupes restants comportait des CFC analogues, mettant en œuvre des relations de synonymie ou d'antonymie. Deux d'entre eux avaient la particularité de rassembler des CFC composées exclusivement de liens de synonymie. L'ensemble des lexies qui les composaient étaient en similarité d'Attributs complète, $nbr\,de\,sim_a = 1 : (nbr\,lexies)^2$. De tels groupes ne sont pas exploitables pour l'enrichissement automatique du RL-fr.

Un seul groupe comportait des CFC, au nombre de deux, qui mettaient en jeu des liens d'antonymie. Elles étaient composées des lexies (OBLIGATOIRE, FACULTATIF, OBLIGATOIREMENT, FACULTATIVEMENT) et ($\text{MALHEUREUX}_{\text{Adj}}$ **II**, $\text{HEUREUX}_{\text{Adj}}$ **II**, MALHEUREUSEMENT, HEUREUSEMENT). Elles relevaient d'une même configuration de dérivation et huit mesures de similarité d'attributs complète avaient été comptabilisées entre leurs lexies.

Les 17 derniers groupes comportaient des CFC qui mettaient toutes en jeu des relations de synonymie. Nous avons observé que quel que soit le nombre de lexies en jeu dans ces CFC le nombre de similarités d'attributs complètes comptabilisées lui était supérieur de deux : $nbr\,de\,sim_a = 1 : nbr\,lexies + 2$. Ainsi, nous pouvions en comptabiliser cinq dans le groupe de deux CFC composées des lexies (LÉCHER, LÉCHAGE, LÈCHEMENT) et (RÂLER, RÂLE, RÂLEMENT).

Cette analyse nous conforte dans l'idée que les relations de synonymie et d'antonymie jouent un rôle distinct des autres relations dans l'organisation du lexique. De la même manière que nous les avons exclues des Attributs des lexies pour établir notre mesure de similarité, nous devons prendre en compte leur présence pour établir les regroupements. Le critère selon lequel deux CFC regroupées par similarité de Relations sont analogues si leurs lexies sont en situation de similarité d'Attributs strictement deux à deux doit donc être affiné.

### 4.4.3 Composantes isolées

La dernière partie de notre analyse a été consacrée aux 14 CFC isolées lors du regroupement par similarité d'Attributs. Chacune de ces CFC mettait en jeu trois sommets et de deux à six arcs. Parmi les 42 sommets répartis dans ces CFC, plus de 70% n'avait fait l'objet d'aucun travail lexicographique. Les autres étaient en cours de traitement. L'ensemble des 46 arcs en présence rendait compte de liens de FL, dont la moitié appartenait à la famille **Syn**.

En 4.3.3, nous avons émis l'hypothèse que l'absence de prise en compte de l'orien-

[14]Cette pratique concerne majoritairement des vocables dont l'unité lexicale de base appartient au champ sémantique de la couleur. Elle est actuellement l'objet d'une réflexion au sein de l'équipe de lexicographes et n'a pour l'instant fait l'objet d'aucune publication.

tation des relations pouvait être à l'origine de mauvais regroupements. L'analyse des CFC isolées nous a permis de vérifier cette hypothèse et d'observer d'autres cas de figure.

Nous nous sommes intéressée aux couples constitués d'une CFC isolée et de chacune des autres CFC avec lesquelles elle avait été regroupée par similarité de Relations. Nous avons constaté que deux critères permettaient d'expliquer l'absence d'analogie : l'orientation des relations en jeu dans les CFC et les Attributs méta-pdd associés aux sommets. Pour une combinaison de critères donnée, nous avons observé un nombre de similarités d'Attributs complètes constant. Le tableau 4.10 résume le résultat de cette analyse.

| méta-pdd | liens | CFC isolées | $sim_a = 1$ |
|---|---|---|---|
| $\neq$ | $\neq$ | 1 | 2/3 |
| ? | $=$ | 1 | 2/3 |
| $\neq$ | $=$ | 8 | 0/3 |
| $=$ | $\neq$ | 4 | 1/3 |

TAB. 4.10 : Répartition des groupes de composantes non analogues

La première combinaison de critères observée correspondait au cas illustré dans la section 4.3.3 et repris par la figure 4.15. L'une des CFC comporte deux noms et un adjectif, tandis que l'autre comporte un nom et deux adjectifs. L'orientation d'un lien de nominalisation et d'un lien d'adjectivation y est inversée. La seconde a été comptabilisée comme CFC isolée.

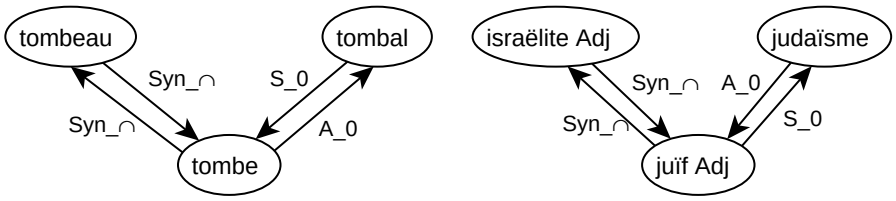



FIG. 4.15 : Exemples de CFC non analogues : méta-pdd $\neq$ liens $\neq$

La seconde combinaison de critères correspondait elle aussi à un seul cas, illustré par la figure 4.16. Elle a permis la détection d'une anomalie dans le réseau. Elle était en effet due à l'absence de partie du discours dans la description de la lexie REMPLACEMENT.

La troisième combinaison de critères correspondait à huit cas de CFC isolées. L'observation de ces CFC nous a montré que la prise en compte de la méta-pdd comme Attribut était un choix pertinent. En effet, elles mettaient toutes en œuvre des FL qui permettent d'encoder des relations entre lexies de catégories variées. La

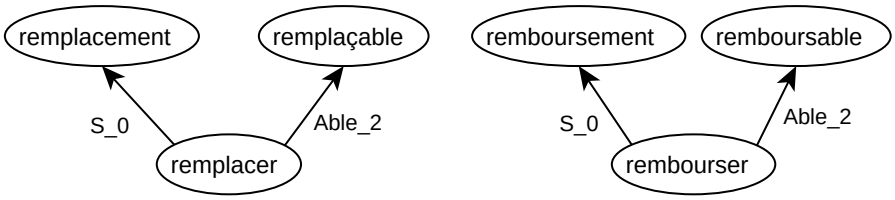



FIG. 4.16 : Exemples de CFC non analogues : méta-pdd ? liens =

figure 4.17 en fournit un exemple. Les relations de dérivation sémantique nominale (**FL31**) et adjectivale (**FL104**) de premier actant qui y sont présentes contraignent uniquement la partie du discours de leur cible. Elles ont ici pour source dans un cas un nom, dans l'autre un verbe. Les CFC ne sont donc pas analogues et se sont retrouvées toutes deux isolées.

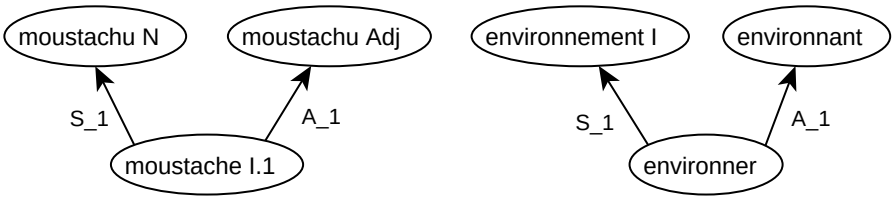



FIG. 4.17 : Exemples de CFC non analogues : méta-pdd ≠ liens =

La quatrième combinaison de critères concernait quatre CFC isolées. Elles avaient toutes la particularité de contenir un cas de liens de FL inverses correctement encodé et un pour lequel l'un des liens était absent. Un lien inverse manquait également aux CFC avec lesquelles elles étaient initialement regroupées. La figure 4.18 montre un exemple de telles CFC. Un lien de verbalisation devrait être ajouté dans celle de gauche — entre les lexies SIMPLIFICATION et SIMPLIFIER — comme dans celle de droite — entre les lexies REPRODUCTIBLE et REPRODUIRE. Ce sont ici les Attributs générés à partir des liens de FL associés à chaque lexie qui ont permis de séparer les CFC qui n'étaient pas analogues. Une mesure de similarité de Relations prenant en compte l'orientation des arcs aurait cependant tout aussi bien fait l'affaire.

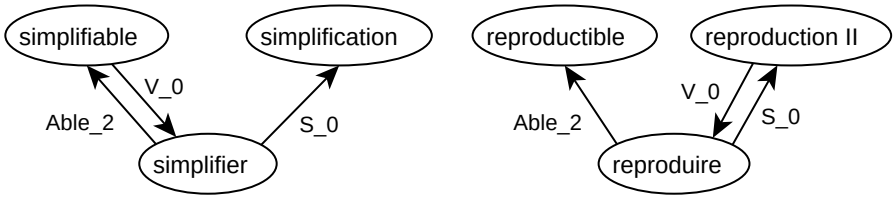



FIG. 4.18 : Exemples de CFC non analogues : méta-pdd = liens ≠

## Conclusion

L'expérience que nous venons de détailler confirme l'hypothèse d'un lexique organisé en sous-groupes d'unités correspondant à des structures de relations récurrentes, identifiables par automatisation du raisonnement analogique. Elle montre également que notre approche fournit des résultats complémentaires à ceux d'une approche de la dérivation lexicale par analogie formelle.

La méthode que nous avons mise au point semble répondre à nos attentes et retourne des groupes de sous-graphes satisfaisants. Elle ne nous a cependant pas permis de vérifier l'ensemble de nos hypothèses. Les lexies présentes dans les composantes faiblement connexes exploitées étaient en effet trop peu décrites. Si elles nous ont permis de vérifier la pertinence des Attributs rendant compte de leur métapdd, celle des Attributs établis à partir des liens de FL qui leur étaient associés n'a pas pu être confirmée. De plus, il ne nous a pas été possible d'établir des configurations de dérivations lexicales comportant à la fois un ensemble de relations et des profils de lexies. La question du développement d'une méthode d'abstraction de telles configurations à partir des regroupements effectués n'a donc pas pu être abordée.

Nous regrettons également de n'avoir pu observer aucune configuration mettant en jeu des relations de copolysémie. Nous espérons que le passage des composantes connexes aux sous-structures moins aisément isolables que sont les motifs locaux nous permettra de remédier à ces insuffisances.

Enfin, cette expérience a mis en avant la nécessité d'une évaluation des configurations de dérivations lexicales par les lexicographes. La faible quantité de données disponible nous a permis de mener ici cette évaluation de manière informelle. Une méthodologie précise devra cependant être mise au point. Elle devra permettre de déterminer des critères de bonne granularité des configurations et d'évaluer les risques de propagation d'erreurs liés à leur exploitation.

# Chapitre 5

# Motifs analogues

## Sommaire

# Introduction

Au cours du chapitre 4, nous avons concentré notre attention sur les agrégats lexicaux de petites tailles. Nous avons émis l'hypothèse que leur exploration par raisonnement analogique permettrait de faire émerger des structures récurrentes à partir desquelles ils seraient possible d'abstraire des configurations de dérivations lexicales. Après une réflexion sur différentes délimitations de sous-graphes[1] possibles comme point d'entrée de cette exploration, nous avons fixé notre attention sur les motifs locaux.

Les implémentations du raisonnement analogique et de la collecte de motifs sont toutes deux réputées coûteuses en temps d'exécution et en mémoire vive. Nous avons donc choisi de tester la mise en place d'une procédure de regroupement de microstructures analogues sur les sous-graphes directement accessibles que représentaient les composantes connexes du RL-fr.

La méthode mise au point a fourni des résultats satisfaisants. Les regroupements que nous avons obtenus étaient cependant trop peu nombreux et composés de données trop faiblement décrites pour valider l'ensemble de nos hypothèses.

Le présent chapitre propose de détailler une nouvelle expérience, réalisée en couplant cette méthode à la collecte de motifs locaux. Après une rapide présentation de la version du RL-fr utilisée, il rappelle la procédure générale en précisant les quelques modifications qui lui ont été apportées. Une analyse des groupes de microstructres obtenus est alors proposée, d'abord sous un angle topologique, puis au regard des choix d'Attributs effectués. La question d'un assouplissement est alors abordée. En dernier lieu, une simulation d'abstraction de configuration de dérivations lexicales est réalisée. Une réflexion sur les prises de décisions impliquée dans ce processus et son exploitation ultérieure est alors entamée.

## 5.1 Un RL-fr sans inclusion formelle

L'expérience sur laquelle nous nous appuyons dans ce chapitre a été réalisée sur une version du RL-fr datant du 6 mai 2014. Comme nous l'avons évoqué au cours du chapitre 4, section 4.4.2, les sous-graphes composés de liens d'inclusion formelle, qui servent à encoder les relations entre les phrasèmes et les lexies qui les composent, sortent du cadre de la dérivation lexicale qui nous intéresse. Nous avons donc pris la décision d'exclure les 4 957 arcs ainsi typés du graphe utilisé pour l'exploration à venir. La tableau 5.1, page suivante, fournit le pedigree du graphe ainsi obtenu.

Nous pouvons constater que cette version du RL-fr était composée de près de 3% de sommets de plus que la version du 30 mars 2014, que nous avions exploitée précédemment. L'observation des descriptions lexicographiques encapsulées dans ces sommets montre peu d'avancées en terme de répartition : 61% d'entre eux n'avaient fait l'objet d'aucun travail, 38% étaient en cours de traitement, 260 en cours de validation et quatre étaient considérés comme disposant d'une description complète. Le

[1] Nous rappelons que cette question est abordée dans la section 4.2.

| | | | |
|---|---|---|---|
| sommets | 22 577 | coefficient d'agrégation | 0,1627 |
| arcs | 41 721 | Distribution des degrés entrants | |
| degré sortant moyen | 1,8476 | a | -2,61509 |
| boucles | 35 | $r^2$ | 0,9201 |
| arcs multiples | 431 | Plus grande composante connexe | |
| arcs symétriques | 21 129 | sommets | 13 886 |
| sommets isolés | 4 563 | arcs | 36 515 |
| composantes faiblement connexes | 5 894 | L | 14,0472 |

TAB. 5.1 : RL-fr sans lien d'inclusion formelle du 6 mai 2014

nombre de descriptions contenant une étiquette sémantique était toujours de près de la moitié et un peu moins d'un tiers ne disposait d'aucun exemple, tandis que 50% en disposait d'un seul. En revanche, il n'en demeurait que 8 ne faisant mention d'aucune CG et près de 40% contenaient une FP.

Les liens d'inclusion formelle ayant été exclus, le nombre d'arcs avait légérement diminué. Il était cependant composé de près de 6,5% de liens de FL de plus et de 74% de liens de copolysémie supplémentaires.

## 5.2 Collecte de microstructures analogues

Le méthode de regroupement de microstructures analogues utilisée ici est très proche de celle présentée dans le chapitre 4. La substitution de sa première étape, le regroupement de structures isomorphes, par la collecte de motifs locaux a cependant nécéssité quelques adaptations.

### 5.2.1 Collecte de sous-graphes

Rapellons-le, le temps d'exécution du dénombrement et de la collecte des motifs d'un graphe dépend à la fois de sa densité, du degré moyen de ses sommets, du nombre de motifs qu'il comporte et du nombre d'occurrences de chacun d'entre eux. De plus, les motifs de grande taille sont plus nombreux que les motifs de petite taille. Nous avons donc choisi de débuter notre exploration par le dénombrement et la collecte de motif de taille 3.

Comme nous l'avons annoncé au chapitre 4, nous avons utilisé pour cela les fonctionnalités dédiées aux motifs de la librairie python *graph-tool* développée par Peixoto (2014). Les occurrences d'un même motifs étant isomorphes, ce traitement vient remplacer la première étape présentée dans la section 4.3.2 du chapitre 4. Cherchant à établir un regroupement de sous-graphes, nous avons choisi de ne collecter que les motifs disposant d'un minimum de deux occurrences. Sur un Mac OS X 10.6.8, muni d'un processeur 3.6 GHz Intel Core i5 disposant d'une mémoire vive de 4 Go 1333 MHz DDR3, cette tâche a nécessité 4 heures, 45 minutes et 15 secondes.

À l'issue de ce traitement, nous disposions de 291 596 sous-graphes répartis en 162 groupes. Cela signifie que 162 motifs apparaissaient plus d'une fois dans le RL-fr sans lien d'inclusion formelle du 6 mai 2014[2].

Seuls 29% de ces groupes de sous-graphes isomorphes ne comportaient que deux occurrences, tandis que près de 10% en comportait plus de mille. La figure 5.1 présente la structure des cinq motifs le plus fréquents avec, de gauche à droite, 118 381, 49 145, 35 205, 27 482 et 16 567 occurrences.

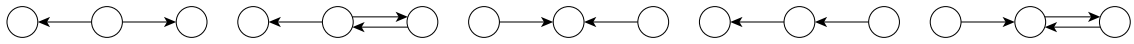


FIG. 5.1 : Motifs 3 les plus fréquents du RL-fr sans lien d'inclusion formelle

Chaque sommet du graphe initial se trouvait impliqué dans différentes occurrences de motifs. Nous avons ainsi observé que les 18 groupes isomorphes ne comportant que deux occurrences, composés de 36 sous-graphes et par conséquent de 108 sommets, ne concernaient que 62 lexies distinctes.

### 5.2.2 Sélection de connexions lexicales

Nous avons ensuite procédé à la comparaison des connexions lexicales en jeu dans chacun des ensembles de sous-graphes isomorphes. Cette étape, réalisée en 16 minutes et 13 secondes, a abouti à la création de 26 394 groupes de 2 à 3 547 sous-graphes. Au total, 278 301 sous-graphes étaient encore présents, correspondant à 139 motifs distincts.

Nous avons alors décidé de réduire cette masse de données. D'une part, nous souhaitions diminuer le temps de traitement nécessaire à la comparaison par similarité de descriptions lexicographiques. D'autre part, nous voulions aboutir à un ensemble de classes analogues qu'il soit possible d'analyser manuellement en détail pour vérifier les hypothèses restées en suspens à l'issue de l'expérience du chapitre 4.

Dans un premier temps, nous avons choisi d'exclure de ces groupes tous ceux qui mettaient en jeu au moins une FL des familles **Syn**, **Anti** et **Contr**, dont nos observations précédentes avaient mis en avant le comportement particulier. Nous avons ainsi obtenu une sous-sélection de 173 508 sous-graphes, correspondant à 80 motifs distincts et répartis en 17 871 groupes. Chacun de ces groupes comportait entre deux et 2 767 sous-graphes.

Dans un second temps, nous avons poursuivi la réduction de notre champ d'observation en écartant l'ensemble des sous-graphes composés de seulement deux arcs. Nous pensions alors que les sous-graphes mettant en jeu davantage de liens de dérivations étaient plus intéressants. De plus, ce critère nous permettait d'aboutir au nombre raisonnable de 5 391 groupes, correspondant à 19 motifs et comportant de

[2]Parallèlement à la collecte décrite ici, nous avons observé que 416 motifs de taille 3 apparaissaient au moins une fois dans cette version du RL-fr privée de liens d'inclusion formelle.

deux à 240 sous-graphes. Nous disposions alors d'un total de 24 967 sous-graphes[3].

### 5.2.3 Comparaison de descriptions lexicographiques

Nous avons vu dans le chapitre 2, section 2.2.2, que la valeur d'application d'une FL est un ensemble et qu'il existe autant de liens de FL que de lexies contenues dans cet ensemble. Ainsi, **Magn**(*aboyer* I) = {*furieusement* I ; *férocement*} correspond à deux liens distincts : ABOYER I → FURIEUSEMENT I et ABOYER I → FÉROCEMENT.

La lexie ABOYER I étant par ailleurs liée à la lexie ABOIEMENT I par des relations de nominalisation et de verbalisation, nous pouvons prévoir que les deux sous-graphes présentés dans la figure 5.2 aient été collectés comme occurrences d'un même motif et conservés par similarité de Relations.

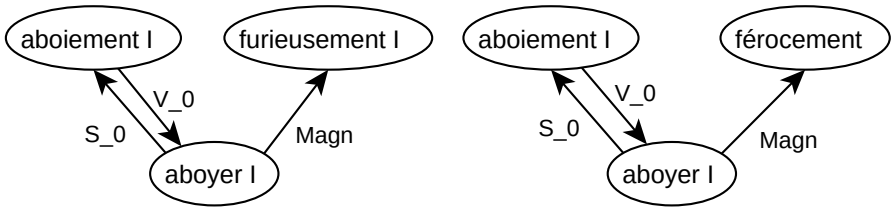



FIG. 5.2 : Sous-graphes comportant deux sommets identiques

Un tel couple de sous-graphes, s'il relève bien d'une même configuration de dérivations lexicales, ne nous apprend rien de plus qu'une simple consultation du RL-fr. Nous avons donc décidé d'ajouter un critère supplémentaire à la procédure de regroupement des sous-graphes par similarité d'Attributs présentée au chapitre 4, section 4.3.4.

Lors de la comparaison de deux sous-graphes, le nombre de lexies distinctes en jeu dans les similarités d'Attributs complètes est comptabilisé. S'il est inférieur à 6, cela signifie soit que les deux sous-graphes partagent des lexies, soit qu'au moins une lexie de l'un des sous-graphes est en jeu dans plus d'une des trois mesures de similarité d'Attributs complète obtenues. Dans les deux cas, le couple n'est pas conservé.

Le reste de la procédure est identique à celle que nous avons précédemment présentée. Chaque sommet est associé à un ensemble d'Attributs, constitué de la manière suivante :

- un Attribut rendant compte de sa méta-pdd ;
- autant d'Attributs FLout que de familles de FL en jeu dans l'ensemble de ses liens de FL sortants ;
- autant d'Attributs FLin que de familles de FL en jeu dans l'ensemble de ses liens de FL entrants ;

[3]Soulignons que ce nombre de sous-graphes correspond à près de 115 fois plus que la quantité obtenue à l'issue de la même étape dans le cadre de l'expérience détaillée dans le chapitre 4.

- un Attribut rendant compte du rapport arithmétique entre le nombre de liens entrants et le nombre de liens sortants, valant **out**+ en cas de supériorité numérique des liens sortants, **in**+ en cas de supériorité numérique des liens entrants ou **in**=**out** en cas d'égalité.

La pertinence du premier de ces Attributs ayant déjà été expérimentée, nous entendons ici vérifier celle des suivants. Il s'agira de déterminer s'ils entrent en jeu dans la distinction de classes analogiques et si les discriminations qu'ils opèrent sont pertinentes.

## 5.3 Groupes de microstructures analogues

À l'issue de ces trois étapes, nous disposions de 124 groupes de microstructures analogues. La répartition des sous-graphes à l'intérieur de ces groupes était assez proche de ce que nous avions observé pour les composantes faiblement connexes. Ainsi, 56% d'entre eux ne comportaient que deux sous-graphes. La dimension du plus grand groupe était cependant supérieure à nos observations précédentes, avec 27 éléments. Au total, 501 sous-graphes avaient été conservés, ce qui représentait moins de 2‰ des motifs collectés.

### 5.3.1 Topologie

Les sous-graphes conservés et regroupés au cours de cette expérience correspondaient à dix motifs parmi les 25 plus fréquents de la collecte initiale. La figure 5.3 propose une représentation sagittale de chacun d'entre eux.

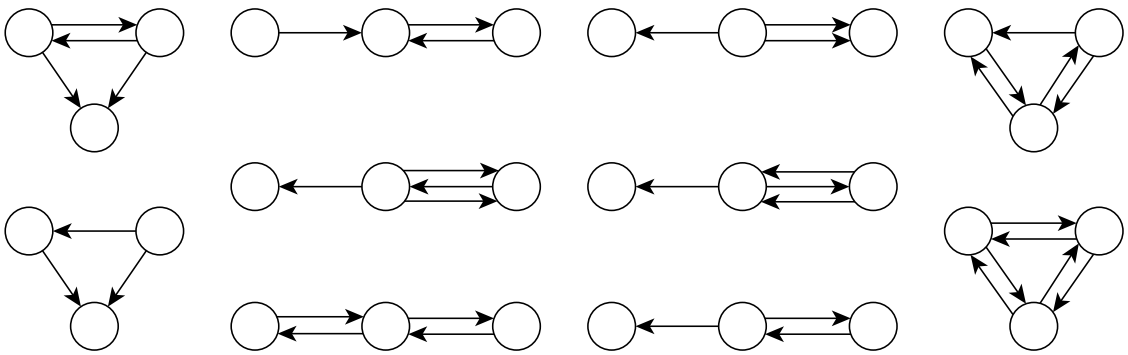


FIG. 5.3 : Motifs en jeu dans les classes de microstructres analogues

Nous pouvons voir sur cette figure que le nombre d'arcs qui composent les sous-graphes varie de trois à six. Au total, nous dénombrions 1 593 arcs. Moins de 4% d'entre eux mettaient en jeu des dérivations sémantiques de copolysémie : par métaphore, métaphore$_{\text{comme si}}$, métonymie et extension. Les autres impliquaient 31 FL distinctes. Le tableau 5.2 présente les cinq plus fréquentes. Nous y retrouvons quatre des FL les plus fréquentes des composantes connexes analogues, accompagnées d'une cinquième, $\mathbf{S_1}$, permettant d'encoder la dérivation sémantique nominale de premier actant.

| Liens | Étiquette | FL | Exemple |
|---|---|---|---|
| 402 | FL21 | $\mathbf{S_0}$ | POÉTIQUE → POÉSIE |
| 267 | FL23 | $\mathbf{V_0}$ | CAMBRIOLAGE → CAMBRIOLER |
| 157 | FL28 | $\mathbf{A_0}$ | POÉSIE → POÉTIQUE |
| 133 | FL29 | $\mathbf{S_1}$ | COMBRIOLER → CAMBRIOLEUR |
| 88 | FL104 | $\mathbf{Adv_0}$ | POÉSIE → POÉTIQUEMENT |

TAB. 5.2 : FL les plus fréquentes dans les occurrences de motifs analogues

Pour leur part, les 1 503 sommets en présence mettaient en jeu 1 079 lexies distinctes. Parmi celles-ci, 852 avait une occurrence unique et la plus fréquente, MÈRE[1] **I.1a** [*Ma mère m'a mis au monde dans une cave.*], apparaissait vingt-deux fois. Un peu plus de 60% de ces lexies n'avaient fait l'objet d'aucun travail lexicographique, 38% étaient en cours de traitement et cinq étaient en cours de vérification. Aucune des quatre lexies disposant d'une description complète n'était impliqué.

### 5.3.2 Impact de la faible description des lexies

La faible avancée des descriptions lexicographiques était notamment visible dans 45 des groupes constitués. Ils avaient en effet la particularité de ne comporter que des sommets dans la description desquels seules les relations en jeu dans la structure de connexions lexicales qui les avaient réunis lors de la seconde étape étaient encodées. La figure 5.4 fournit un exemple de connexions lexicales et d'ensembles d'Attributs en jeu dans un tel groupe.

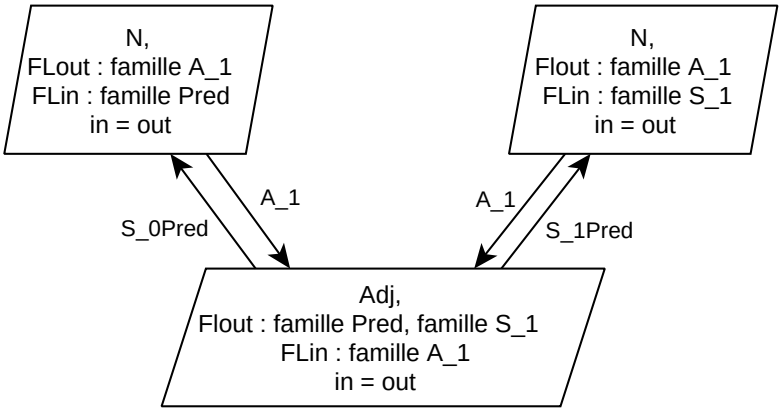



FIG. 5.4 : Connexions lexicales entre lexies peu décrites

Cet exemple est tiré d'un groupe de deux sous-graphes, permettant d'énoncer trois classes de proportions analogiques équivalentes.

ORIGINALITÉ : $\text{ORIGINAL}_{\text{N}}$ **II** :: SOLIDITÉ **II** : $\text{SOLIDE}_{\text{N}}$,
ORIGINALITÉ : $\text{ORIGINAL}_{\text{Adj}}$ **II** :: SOLIDITÉ **II** : $\text{SOLIDE}_{\text{Adj}}$ **I**,

ORIGINAL$_{\text{N}}$ **II** : SOLIDE$_{\text{N}}$ **II** :: SOLIDITÉ **II** : SOLIDE$_{\text{Adj}}$ **I**.

Bien que correctes, ces proportions analogiques sont très peu informatives. Les lexies qu'elles impliquent ne disposaient alors ni d'étiquette sémantique, ni de FP, pas plus que d'exemple lexicographique. Il aurait donc été prématuré de les exploiter pour établir une configuration de dérivations lexicales.

Nous avons observé que 16 des 45 groupes partageant cette particularité entraient en concurrence avec au moins un autre groupe de sous-graphes analogues mettant en jeu les mêmes Relations, mais dont les sommets étaient davantage décrits. Le regroupement comportant les lexies les moins décrites était alors systématiquement le plus étoffé.

## 5.4 Pertinence des Attributs sélectionnés

Les seize groupes de sous-graphes composés de lexies peu décrites que nous venons d'évoquer ne sont pas les seuls à entrer en concurrence avec au moins un autre groupe mettant en jeu les mêmes connexions lexicales. Au total, nous en dénombrons 77, issus de 23 ensembles de connexions lexicales distincts.

Nous avons mené une analyse détaillée de ces groupes, afin de déterminer selon quels critères ils avaient été discriminés. Nous avons ainsi pu déterminer la pertinence des Attributs laissés pour compte lors de l'analyse des composantes faiblement connexes analogues.

### 5.4.1 Scissions artificielles

La première de nos observations a été que six groupes de connexions lexicales identiques avaient été scindés artificiellement en quatorze groupes de sous-graphes analogues. Pour l'ensemble de ces cas, le critère du nombre minimal de lexies distinctes en jeu dans les similarités d'Attributs complètes était en cause.

La figure 5.5 fournit un exemple d'un tel ensemble de connexions lexicales et d'Attributs de sommets communs à deux groupes de microstructures analogues.

Le premier de ces groupes mettait en jeu les triplets de lexies[4] (BOUCHER$_{\text{N}}$ **I.1**, BOUCHERIE **I**, MARCHAND$_{\text{N}}$) et (BOULANGÈRE **2**, BOULANGERIE **I**, MARCHANDE), le second les triplets (BOUCHÈRE **1**, BOUCHERIE **I**, MARCHANDE) et (BOULANGER$_{\text{N}}$ **2**, BOULANGERIE **I**, MARCHAND$_{\text{N}}$).

La séparation de ces triplets n'était pas souhaitable et nous aurions préféré être en mesure d'énoncer des proportions analogiques telles que « BOULANGÈRE **2** est à BOULANGERIE **I** ce que BOUCHÈRE **1** est à BOUCHERIE **I** ».

[4]La lexie BOULANGÈRE **2** désigne la vendeuse d'une boulangerie. Elle est dérivée sémantiquement d'une première acception du vocable BOULANGÈRE, qui désigne la personne qui fabrique le pain et les viennoiseries.

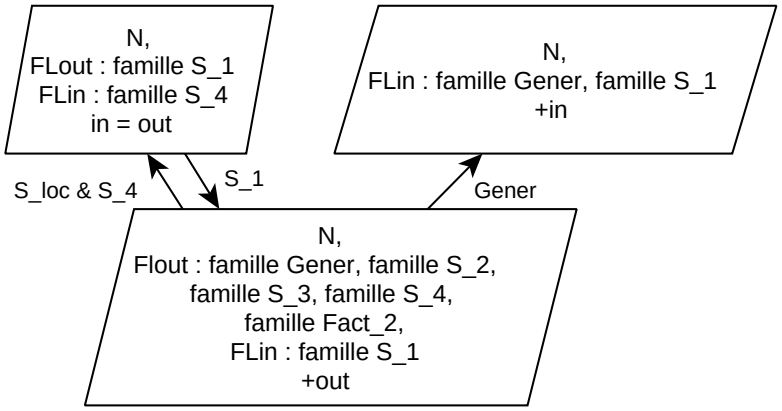



FIG. 5.5 : Connexions lexicales et Attributs de sommets communs à deux groupes de microstructures analogues

Nous avons également remarqué que quatre des groupes ainsi constitués n'étaient pas composés de sous-graphes analogues. En effet, une partie des sommets entre les descriptions desquels une similarité d'Attributs complète avait été mesurée n'occupait pas la même position dans la structure de ces sous-graphes.

La figure 5.6 fournit une illustration de l'un de ces groupes. Nous pouvons y voir que MÈRE[1] **I.1a** [*Ma mère m'a mis au monde dans une cave.*] et PÈRE **I.1a** [*Il est père de deux enfants.*] d'une part, et FILLE **II** [*Je viens de fêter l'anniversaire de ma fille.*] et FILS **I** [*Le fils de Flora aura bientôt six mois.*] d'autre part, n'y jouaient pas le même rôle.

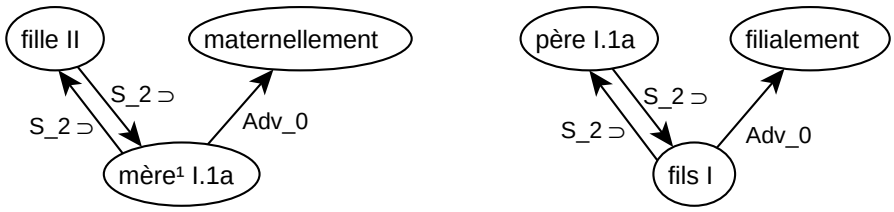



FIG. 5.6 : Erreur de regroupement de microstructures

Malgré ces considérations, nous n'envisageons pas de supprimer la contrainte du nombre minimal de lexies distinctes en jeu dans les similarités d'Attributs complètes. Nous persistons à penser qu'elle permet d'écarter des regroupements peu pertinents[5]. Il serait alors préférable de mettre au point un post-traitement fusionnant les groupes entre lesquels aucune distinction d'Attributs ni de Relations ne peut être observée.

### 5.4.2 Familles de FL

Les 63 autres groupes concurrents, issus de 17 ensembles de connexions lexicales distincts, avaient tous été discriminés à partir des Attributs que nous avions choisi

[5]Cette possibilité a été testée dans le cadre d'une seconde expériences qui sera évoquée dans la conclusion de ce chapitre.

de comparer.

Le seul critère de différenciation à l'œuvre dans la création de 48 de ces groupes, correspondant à 11 ensembles de connexions lexicales, était la présence ou l'absence de certaines familles de FL dans les liens entrants et sortants des sommets qu'ils impliquaient. Ainsi, parmi les ensembles d'Attribtus présentés dans la figure 5.7, nous pouvions observer une variation des familles de FL pour les liens entrants des lexies adjectivales.

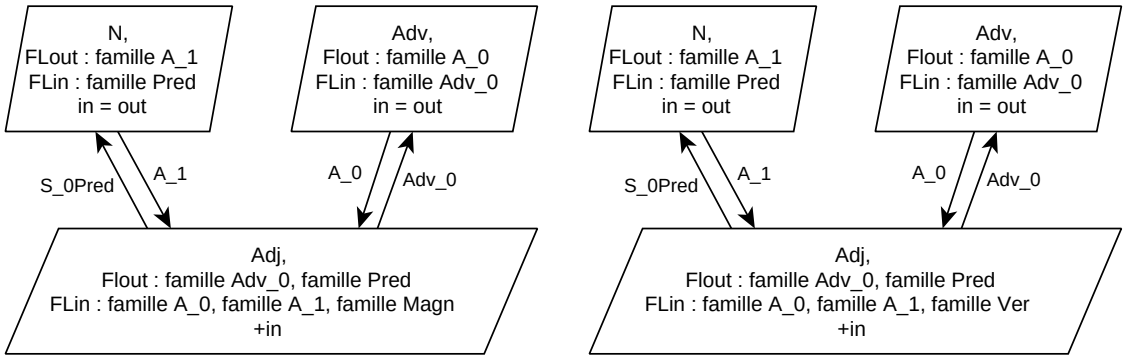



FIG. 5.7 : Variation de familles de FL en jeu

Un premier groupe de sous-graphes analogues, correspondant à la répartition d'Attributs de gauche, mettait en jeu les triplets de lexies (INTENSE, INTENSITÉ **I**, INTENSÉMENT **1**) et (PRÉCOCE, PRÉCOCITÉ, PRÉCOCEMENT), dont les lexies adjectivales étaient encodées en tant que modificateurs d'intensification, famille **Magn**, d'au moins une autre lexie du RL-fr. Cette relation syntagmatique correspond notamment à celle en œuvre entre CHALEUR **I.1b** et INTENSE dans l'énoncé *La chaleur intense stimule la circulation sanguine et accélère l'élimination des toxines*[6].

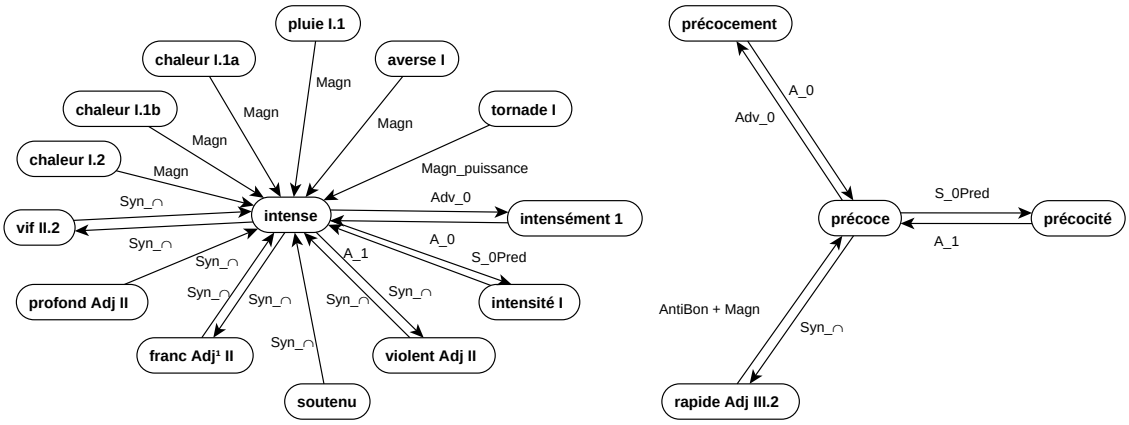



FIG. 5.8 : Variation de liens de FL

[6]Cet exemple est extrait du FrWac qui le référence comme étant issu du nom de domaine `lycoseshop.fr`

Comme nous pouvons le voir sur la figure 5.8, un tel rapprochement n'aurait pas été possible si nous avions comparé directement les liens de FL en jeu, sans une factorisation en familles. La variation du nombre de liens en jeu, la présence d'un lien de FL **Magn$_{\text{puissance}}$** et l'absence de lien de FL **AntiBon + Magn** dans la connectivité de la lexie INTENSE n'auraient alors pas permis de la considérer comme étant similaire à la lexie PRÉCOCE.

Le second groupe mettait en jeu les triplets de lexies (OBJECTIF$_{\text{Adj}}$, OBJECTIVITÉ, OBJECTIVEMENT), (PRÉCIS$_{\text{Adj}}$, PRÉCISION, PRÉCISÉMENT) et (PERTINENT, PERTINENCE, PERTINEMMENT), dont les lexies adjectivales étaient encodées en tant que modificateurs de « confirmation », famille **Ver**, d'au moins une autre lexie dans le RL-fr. Cette relation syntagmatique correspond notamment à celle en œuvre entre CHIFFRE **I.2** et PRÉCIS$_{\text{Adj}}$ dans l'énoncé *S'il n'est pas possible de donner un chiffre précis, cela implique que les chiffres sont faux*[7].

La distinction opérée ici nous semble pertinente. Si l'énonciation d'une analogie telle que « INTENSE est à INTENSITÉ **I** ce que PERTINENT est à PERTINENCE » n'est pas incorrecte, celle d'une analogie telle que « PERTINENT est à PERTINENCE ce que PRÉCIS$_{\text{Adj}}$ est à PRÉCISION » nous paraît plus fine et plus informative. Pour la première, le rapport pourrait être énoncé à l'aide de la seule FL encodée entre les couples de lexie : *un dérivé sémantique adjectival de premier actant*. Pour la seconde, l'information supplémentaire encodée à l'aide de la famille de FL **Ver** pourrait être ajoutée : *un dérivé sémantique adjectival de premier actant pouvant être employé comme modificateur de « confirmation »*.

Les autres distinctions opérées sur ce même critère unique nous ont toutes semblé aussi intéressantes. Les familles de FL en jeu dans les liens entrants et sortants des sommets lexicaux permettaient donc bien de contraindre suffisamment la nature des objets entre lesquels nous cherchons à établir une conformité de rapports pour aboutir à l'énonciation d'analogies pertinentes.

### 5.4.3 Rapport arithmétique entre liens entrants et sortants

Nous n'avons pu observer qu'un seul cas dans lequel la différenciation entre deux groupes était réalisée sur le seul critère d'une variation du rapport arithmétique entre les nombres de liens de FL entrants et sortants des lexies impliquées dans leurs sous-graphes. Il est illustré dans la figure 5.9.

Le groupe correspondant aux attributs de la structure de gauche comportait deux sous-graphe, composés des lexies (FINIR **I.1**, FIN$_{\text{N}}$ **I.1**, FINISSANT) et (APPUYER **II**, APPUI **I.1**, APPUYÉ). Les lexies nominales qui y apparaissaient avaient la particularité de compter deux liens de FL sortant typés $\mathbf{V_0}$ dans le RL-fr, pour un seul lien entrant, typé $\mathbf{S_0}$. Elles disposaient donc d'un Attribut +**out**. Nous pouvions également noter que, dans les deux cas, l'un des liens $\mathbf{V_0}$ avait pour cible un verbe pronominal.

Les lexies nominales du second groupe, en revanche, ne comptaient qu'un seul

[7]Cet exemple est extrait du FrWac qui le référence comme étant issu du nom de domaine `barfumeur.forumpro.fr`

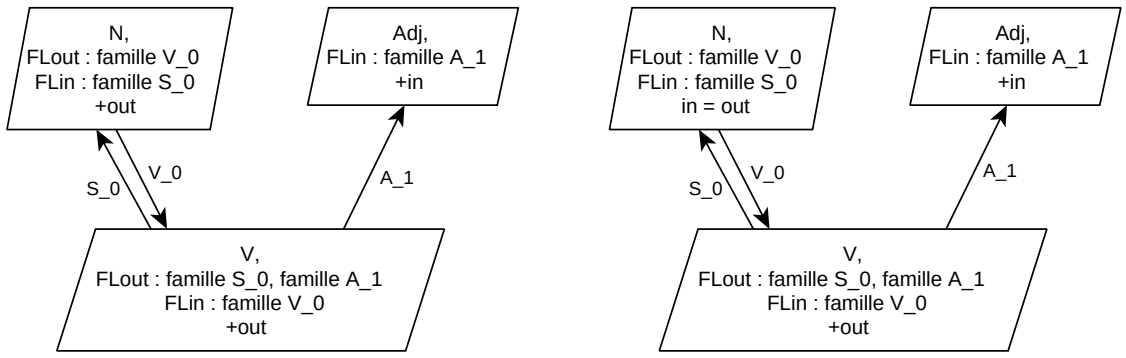



FIG. 5.9 : Variation de rapport arithmétique entre liens entrants et sortants

lien de FL sortant, typé $\mathbf{V_0}$ et un seul lien de FL entrant, typé $\mathbf{S_0}$. Elles disposaient donc d'un Attribut **in=out**. Parmi les seize sous-graphes rassemblaient dans ce groupe, un seul impliquait un verbe pronominal : (S'ORIENTER **I**, ORIENTATION **I**, ORIENTÉ **I**).

Bien que la distinction opérée ici ne mettent en jeu que des lexies peu décrites, elle ne nous paraît pas dénuée de sens. Nous regrettons cependant de ne pas avoir eu l'occasion d'observer des cas plus emblématiques, opposants des lexies carrefours telles que FAIRE **II.1** [*Il fait du ping-pong.*] à des lexies plus « classiques », telles que JOUER **II.2** [*Il joue au ping-pong*][8].

### 5.4.4 Variation d'ordonnancement des Relations

Tout comme nous avions pu l'observer dans les regroupements de composantes faiblement connexes analogues, certaines comparaisons d'Attributs ont ici permis de distinguer des sous-graphes partageant les mêmes Relations, organisées différemment.

Ainsi, quatre sous-graphes partageant les mêmes connexions lexicales avaient étaient séparés en deux groupes de microstructures analogues à l'aide des Attributs de méta-pdd. Ils sont illustrés dans la figure 5.10. Celui de gauche comportaient les triplets de lexies (PASSÉ$_{\text{Adj}}$ **I.1**, PASSÉ$_{\text{Adj}}$ **I.2**, PASSÉ$_{\text{N}}$) et (FUTUR$_{\text{Adj}}$ **I.1**, FUTUR$_{\text{Adj}}$ **II**, FUTUR$_{\text{N}}$), celui de droite les triplets (BOUCHE **I.1a**, BOUCHE **I.3**, BUCAL) et (CIL **I.1a**, CIL **II**, CILAIRE).

Deux autres groupes, partageant un autre ensemble de connexions lexicales suivant une organisation différente, avaient été établis à partir de la combinaison de l'ensemble des types d'Attributs disponibles. Ils sont illustrés dans la figure 5.11. Celui de gauche rassemblait les triplets de lexies (BELGE$_{\text{Adj}}$ **1**, BELGE$_{\text{Adj}}$ **2**, BELGIQUE) et (ALLEMAND$_{\text{Adj}}$ **1**, ALLEMAND$_{\text{Adj}}$ **2**, ALLEMAGNE), celui de droite en comportait sept, tous composés de lexies peu décrites. Nous pouvions notamment y observer les triplets (FÊTE **1**, FÊTE **2**, FESTIF) et (DIMENSION **1**, DIMENSION **2**, DIMENSIONNEL).

[8]Nous n'avons pas non plus observé de cas d'exclusion par similarité d'Attributs basée sur ce critère parmi les sous-graphes partageant les mêmes connexions lexicales que ceux des 124 groupes analogues.

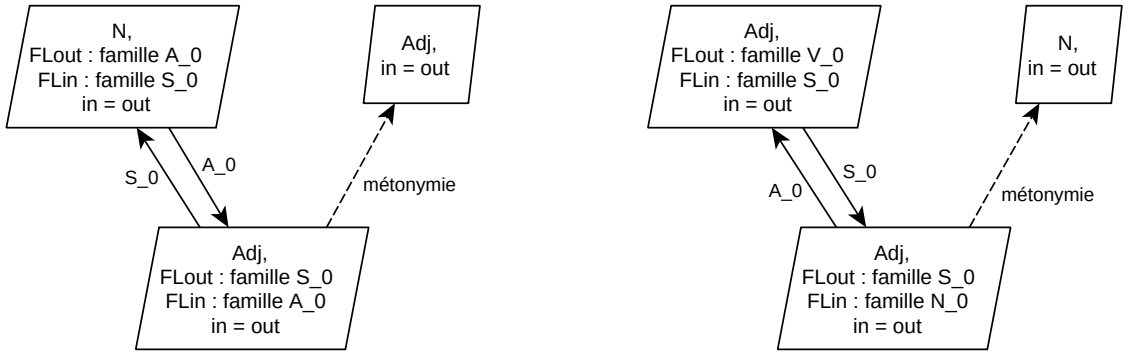



FIG. 5.10 : Variation de méta-pdd

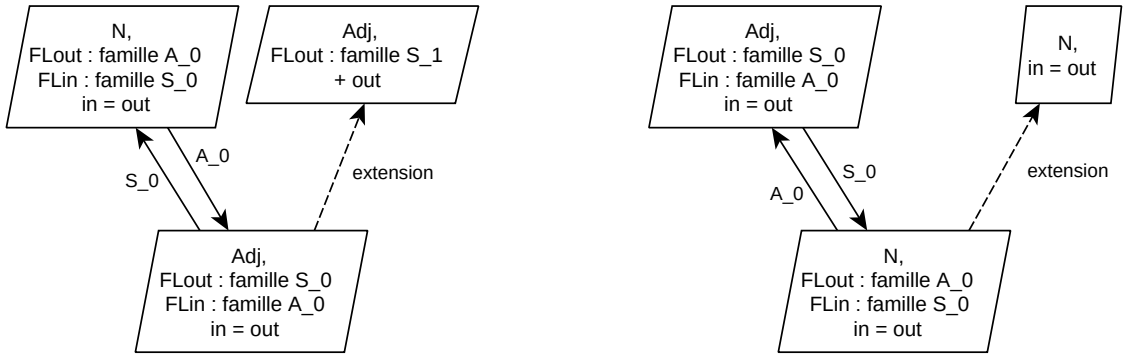



FIG. 5.11 : Variation de l'ensemble des Attributs

Enfin, deux groupes, partageant un troisième ensemble de connexions lexicales orientées différemment, avaient été distingués par la combinaison de leurs Attributs de familles de FL et de rapports arithmétiques entre le nombre de liens entrants et sortants. Ils sont illustrés par la figure 5.12. Celui de gauche contenait trois sous-graphes, parmi lesquels (PIQUE-NIQUE, PIQUE-NIQUER, PIQUE-NIQUEUSE) et (TRAFIC[1], TRAFIQUER, TRAFIQUANT), celui de droite en rassemblait 27, parmi lesquels (GUET, GUETTER, GUETTEUR) et (RONFLEMENT, RONFLER, RONFLEUSE).

Les sous-graphes appartenant aux groupes concurrents dont nous n'avons pas encore parlé disposaient de connexions lexicales ordonnées de manière semblable. Ils avaient tous été répartis par une combinaison d'Attributs de familles de FL et de rapports arithmétiques entre le nombre de liens entrants et sortants de leurs sommets. Dans tous les cas observés, l'un des regroupements permettait d'établir des analogies plus fines que le second, impliquant des lexies peu décrites. Ainsi, les triplets de lexies (DÉFENDRE**I**, DÉFENSE, DÉFENDABLE) et (LIBÉRER[I], LIBÉRATION, LIBÉRABLE) avaient été séparés de triplets aussi variés que (TOURNER**II.1**, TOURNEMENT, $\text{TOURNANT}_{\text{Adj}}$) et (CROIRE**1**, CROYANCE, CROYABLE). Nous considérions alors l'ensemble des types d'Attributs que nous avions selectionné comme étant pertinent.

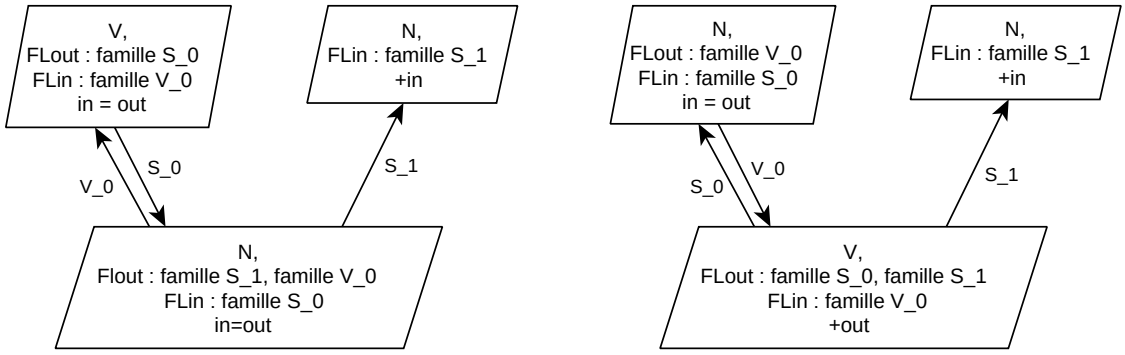



FIG. 5.12 : Organisation différente sans distinction de méta-pdd

## 5.5 Un assouplissement est-il souhaitable ?

Afin de compléter cette analyse, nous nous sommes intéressée aux sous-graphes initialement associés à certains ensembles de connexions lexicales qui n'avaient pas été conservés lors de l'étape de confrontation des descriptions lexicographiques encapsulées dans leurs sommets. Nous souhaitions ainsi déterminer si le critère d'une stricte similarité d'Attribtus complète entre sommets devait être assoupli et si les Attributs sélectionnés engendraient l'exclusion de sous-graphes analogues.

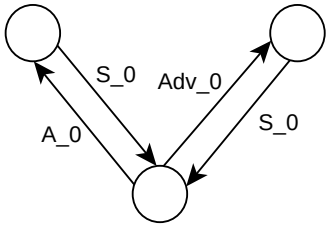



FIG. 5.13 : Connexions lexicales communes à 27 sous-graphes

Nous avons ainsi constaté que parmi les 27 occurrences de motifs partageant l'ensemble de Relations présenté dans la figure 5.13, seules onze avaient été conservées par similarité d'Attributs. Elles avaient été réparties en deux groupes, l'un contenant les triplets de lexies (ART**I**, ARTISTIQUE, ARTISTIQUEMENT) et (PHILOSOPHIE**I**, PHILOSOPHIQUE, PHILOSOPHIQUEMENT), l'autre neuf triplets aussi variés que (NUIT, NOCTURNE, NUITAMMENT) et (INSTINCT, INSTINCTIF, INSTINCTIVEMENT).

Les sous-graphes exclus ne nous semblaient, de prime abord, pas spécifiquement différents de ceux de ces groupes. L'un d'entre eux semblait même très proche de (NUIT, NOCTURNE, NUITAMMENT), puisqu'il était composé des lexies MATIN, MATINAL et MATINALEMENT.

Dans un premier temps, nous avons regardé les similarités d'Attributs partagées par les sommets de chacun des deux groupes de sous-graphes conservés. Rappelons-le, cette similarité est mesurée de la manière suivante :

$$sim_a(Lexie_1, Lexie_2) = \frac{2 \times nbr\ d'attributs\ communs}{nbr\ d'attributs\ Lexie_1 + nbr\ d'attributs\ Lexie_2}$$

Nous avons alors observé que les lexies adjectivales et adverbiales présentes étaient en similarité complète, tandis que les lexies nominales partageaient une similarité inférieure à 0,75 :

$sim_a(\text{ARTISTIQUE}, \text{NOCTURNE}) = 1,$
$sim_a(\text{ARTISTIQUEMENT}, \text{NUITAMMENT}) = 1,$
$sim_a(\text{ART I}, \text{NUIT}) \simeq 0,73.$

Nous nous sommes ensuite intéressée aux similarités d'Attributs des couples formés à partir de l'ensemble d'Attributs de ART I ou de NUIT et de celui de la lexie nominale de chacun des seize triplets écartés. Les mesures ainsi obtenues sont présentées dans le tableau 5.3.

| $sim_a(N, N)$ | ART I | NUIT | $sim_a(N, N)$ | ART I | NUIT |
|---|---|---|---|---|---|
| ALPHABET | 0,62 | **0,83** | MATIN | **0,83** | **0,80** |
| ANARCHIE | **0,83** | 0,73 | MÉDECINE | **0,83** | 0,57 |
| DIEU I | **0,77** | 0,67 | MOIS 1 | 0,67 | 0,57 |
| FAMILLE I.a | 0,71 | 0,50 | NOMBRE I | 0,53 | 0,57 |
| FAMILLE I.b | 0,71 | 0,50 | PROFESSION | **0,77** | 0,67 |
| FORME I.2 | 0,56 | 0,47 | VOIX I.1 | 0,57 | 0,62 |
| GRAMMAIRE a | 0,73 | 0,73 | EXPÉRIENCE II | **0,92** | 0,67 |
| HOMME 1a | 0,62 | **0,83** | NOMBRE II.1 | 0,59 | 0,50 |

TAB. 5.3 : Similarités d'Attributs entre lexies nominales (1)

Nous pouvons y voir que neuf de ces comparaisons aboutissaient à un score supérieur à 0,75, parmi lesquels deux impliquaient la même lexie : MATIN. L'ensemble d'Attributs associé à cette dernière apparaissait toutefois comme étant davantage similaire à celui associé à ART I qu'à celui de NUIT.

Nous avons alors effectué le même type de comparaisons entre les lexies adverbiales d'une part et adjectivales de l'autre. Les résultats obtenus étaient tous égaux à 1, à l'exception des deux suivants :

$sim_a(\text{ARTISTIQUE}, \text{PROFESSIONNEL}_{Adj}\text{I}) \simeq 0,67,$
$sim_a(\text{NOCTURNE}, \text{PROFESSIONNEL}_{Adj}\text{I}) \simeq 0,67.$

Nous aurions ainsi pu résumer la situation en considérant que si nous avions choisi de tolérer une dissimilarité de 25% entre les Attributs des sommets, cinq

sous-graphes suppplémentaires auraient été intégrés au premier groupe obtenu sur la base d'une similarité complète et deux au second.

À ce stade, nous ne nous étions cependant pas assurée de la similarité d'Attributs entre les lexies nominales des sous-graphes ainsi ajoutés. La matrice symétrique présentée dans le tableau 5.4 montre qu'elle est inférieure à 0,75 pour au moins un couple mettant en jeu chacune des lexies assimilables à ART **I**.

| $sim_a(N, N)$ | ANARCHIE | DIEU **I** | EXPÉRIENCE | MATIN | MÉDECINE |
|---|---|---|---|---|---|
| ANARCHIE | 1 | 0,77 | 0,77 | 0,91 | 0,67 |
| DIEU **I** | 0,77 | 1 | 0,71 | 0,83 | 0,88 |
| EXPÉRIENCE | 0,77 | 0,71 | 1 | 0,83 | 0,75 |
| MATIN | 0,91 | 0,83 | 0,83 | 1 | 0,71 |
| MÉDECINE | 0,67 | 0,88 | 0,75 | 0,71 | 1 |

TAB. 5.4 : Similarités d'Attributs entre lexies nominales (2)

Il en va de même pour celles assimilables à NUIT, pour lesquelles la similarité d'Attributs entre les descriptions de ALPHABET et de HOMME **1a** est égale à 0,71.

Dans un dernier temps, nous nous sommes intéressée aux raisons des dissimilarités établies entre ces lexies nominales. La majeure partie d'entre elles se basait sur des différences de familles de FL et était pertinente. Par exemple, MÉDECINE se distinguait de PHILOSOPHIE par le fait que sa description contenait des liens de dérivations sémantiques nominales de deuxième, troisième et quatrième actants le mettant notamment en relation avec PATIENTE, MALADIE **I.1a** et MÉDICAMENT.

Un cas, cependant, était discutable. Il concernait le couple (NUIT,MATIN). Nous avons constaté que le seul Attribut qui permettait de distinguer ces lexies était le rapport arithmétique entre leur nombre de liens entrants et sortants, **in=out** pour la première, **+out** pour la seconde. En consultant leurs descriptions dans l'éditeur Dicet, il nous est apparu que l'unique raison de cette différence était la présence d'un lien d'ajectivation $\mathbf{A_0}$ supplémentaire encodé entre MATIN et MATUTINAL. Cette association était accompagnée de la mention **vieilli litt**.

Au cours de notre travail, nous ne nous sommes pas intéressée aux mentions qui accompagnent les valeurs d'application des FL. Celle-ci nous informe que le lien encodé renvoie à un emploi vieilli et littéraire. Elle rend ainsi compte de certaines CG de la cible du lien, ses marques d'usage. La prise en compte d'une telle information nous aurait cependant permis d'améliorer notre traitement. Nous aurions ainsi pu choisir d'exclure les liens rendant compte de dérivations rares ou vieillies.

Ces observations nous amènent à la conclusion qu'une stricte similarité d'Attributs complète entre les descriptions lexicographiques encapsulées dans les sous-graphes lexicaux doit être conservée. Ainsi, comme nous l'avons établi au chapitre 4, section 4.3.5, les Attributs comparés doivent être considérés comme le socle minimal

commun pour que deux lexies occupent la même place dans une configuration de dérivations lexicales et assurer la pertinence de cette dernière. Nous pensons donc que si des améliorations doivent être introduites, ce n'est pas sur la procédure générale qu'elles doivent porter, mais sur la définition des Attributs comparés.

Ces conclusions sont confirmées par l'observation des conséquences de l'abandon d'une stricte similarité d'Attributs sur certaines distinctions de groupes de sous-graphes que nous avions jugées pertinentes. Ainsi, les groupes partageant les mêmes connexions lexicales et impliquant l'un les triplets (CANCER, CANCÉREUX$_{\text{Adj}}$[1], CANCÉREUSE) et (GRIPPE, GRIPPAL, GRIPPÉ$_{\text{N}}$), le second un ensemble de six triplets impliquant tous des toponymes, à la manière de (ASIE, ASIATIQUE **1**, ASIATIQUE$_{\text{N,fém}}$) et le troisième les triplets (ÉGLISE **II**, ÉCCLÉSIAL, ÉCCLÉSIASTIQUE$_{\text{N}}$) et (NATIONALISME, NATIONALISTE$_{\text{Adj}}$, NATIONALISTE$_{\text{N,masc}}$) auraient tous trois été fusionnés.

## 5.6 Abstraction de configurations de dérivations lexicales

Les lexies en jeu dans les sous-graphes analogues obtenus lors de cette expérience sont encore majoritairement peu décrites. Cependant, certains regroupement nous ont permis d'amorcer une réflexion sur la méthode d'abstraction de configurations de dérivations lexicales que nous souhaiterions mettre en place à l'avenir.

Nous avons distingué trois cas de figure. Le premier était un cas idéal pour lequel l'ensemble des éléments de description encapsulés dans certains sommets de sous-graphes analogues étaient entièrement compatibles. Dans le second cas, ces éléments étaient complémentaires. Dans le troisième, en revanche, certains éléments entraient en contradiction.

Le groupe de six sous-graphes mettant tous en jeu des toponymes, évoqué dans la section précédente, présente la particularité de rassembler ces trois cas de figure. Nous nous appuierons donc dessus pour illustrer notre propos. Le tableau 5.5 présente l'ensemble des triplet de lexies qu'il rassemble.

| AFRIQUE | AFRICAIN$_{\text{Adj}}$ | AFRICAIN$_{\text{N}}$ |
|---|---|---|
| ⌜AMÉRIQUE DU NORD **II**⌝ | AMÉRICAIN$_{\text{Adj}}$ **II.1** | AMÉRICAIN$_{\text{N}}$ **I.2** |
| ⌜AMÉRIQUE DU SUD⌝ | SUD-AMÉRICAIN$_{\text{Adj}}$ | SUD-AMÉRICAIN$_{\text{N}}$ |
| ASIE | ASIATIQUE$_{\text{Adj}}$ **1** | ASIATIQUE$_{\text{N,masc}}$ |
| AUSTRALIE **II** | AUSTRALIEN$_{\text{Adj}}$ **II** | AUSTRALIEN$_{\text{N}}$ |
| EUROPE **I** | EUROPÉEN$_{\text{Adj}}$ **I.1** | EUROPÉEN$_{\text{N}}$ |

TAB. 5.5 : Triplets toponymiques analogues

Nous allons à présent étudier les lexies de ce tableau, colonne après colonne. Nous chercherons ainsi à mettre au point un profil de lexie pour chacune d'entre elles. Ces profils seront ensuite intégrés au squelette de connexions lexicales qui structure les

sous-graphes en jeu, afin d'établir une première configuration de dérivations lexicales toponymiques.

Avant cela, rappellons que, comme nous l'avons détaillé au cours du chapitre 2 section 2.2.4, les éléments des descriptions lexicographiques du RL-fr disposent d'indices de confiance. Certains de ces éléments peuvent donc être considérés comme plus fiables que d'autres. Ainsi, les CG disposent toutes d'un indice de confiance maximal alors que les étiquettes sémantiques sont majoritairement considérées comme incertaines[9]. Certaines décisions concernant les éléments de description devant figurer dans les configurations seront donc moins problématiques que d'autres.

### 5.6.1 Lieu géographique ou entité sociale ?

Les descriptions des lexies de la première colonne partagent toutes les CG `nom propre` et `fém`. Dans le cas de ⌜AMÉRIQUE DU NORD II⌝ et ⌜AMÉRIQUE DU SUD⌝ la caractéristique `locution nominale` est également présente. Nos connaissances du lexique nous permettent de considérer que seule la première de ces trois caractéristiques doit être intégrée au profil de lexie occupant cette position dans la configuration de dérivation lexicale.

D'un point de vue pratique, nous pouvons isoler cette caractéristique à l'aide de la hiérachie implémentée dans la base de données du RL-fr, où elle est enregistrée en tant que partie du discours. Une autre solution possible consiste à laisser de côté les CG dans le cadre de la constitution de profils de lexie et à leur substituer la méta-pdd utilisée lors du regroupement des microstructures analogues. Nous choisissons ici d'opter pour cette dernière solution. Nous verrons par la suite si ce choix est toujours pertinent dans le cas des deux autres colonnes.

Le tableau 5.6 présente les autres éléments de description lexicographique disponibles pour chacun des toponymes. Les exemples lexicographiques, au nombre total de onze, ne sont pas exploités dans le cadre de cette expérience.

Comme nous pouvons le voir dans ce tableau, deux étiquettes sémantiques entrent en concurrence pour les toponymes en présence. Les lexies AFRIQUE, ⌜AMÉRIQUE DU SUD⌝, ASIE et EUROPE I sont associées à l'étiquette lieu géographique avec un indice de confiance de 100%, tandis que les lexies ⌜AMÉRIQUE DU NORD II⌝ et AUSTRALIE II sont associées à l'étiquette entité sociale, l'une avec un indice de confiance de 100%, l'autre de 60%. Cette différence ne semble pas relever d'erreurs d'étiquetage ou d'un mauvais regroupement analogique. Au contraire, elle met en avant une diversité intéressante à mentionner. Elle est accompagnée d'une variation régulière de la valeur d'application de la FL **Gener**. Pour les quatre toponymes étiquetés lieu géographique, cette valeur correspond à la lexie CONTINENT **I.1**. Pour les deux toponymes étiquetés entité sociale, elle correspond à la lexie PAYS **I.1**.

[9]La tâche d'étiquetage sémantique étant particulièrement délicate, une nouvelle consigne a été donnée aux lexicographes le 8 octobre 2014. À compter de cette date, seules les deux lexicographes spécialisées dans la gestion de l'ontologie d'étiquettes sémantiques sont autorisées à attribuer un indice de confiance de 100% à une association lexie-étiquette.

| Lexies | Étiquettes sémantiques | FP | FL |
|---|---|---|---|
| AFRIQUE | lieu géographique | ∼ où vivent les X=1 | **Syn**$_\cap$, **Gener**, **A**$_\mathbf{0}$, **S**$_\mathbf{1}$ |
| ⌜AMÉRIQUE DU NORD ‖⌝ | entité sociale | ∼ où vivent les X=1 | **Syn**, **Gener**, **A**$_\mathbf{0}$, **S**$_\mathbf{1}$ |
| ⌜AMÉRIQUE DU SUD⌝ | lieu géographique | ∼ où vivent les X=1 | **Gener**, **A**$_\mathbf{0}$, **S**$_\mathbf{1}$ |
| ASIE | lieu géographique | ∼ où vivent les X=1 | **Gener**, **A**$_\mathbf{0}$, **S**$_\mathbf{1}$ |
| AUSTRALIE ‖ | entité sociale | ∼ où vivent les X=1 | **Gener**, **A**$_\mathbf{0}$, **S**$_\mathbf{1}$ |
| EUROPE \| | lieu géographique | ∼ où vivent les X=1 | **Syn**$_\cap$, **Gener**, **A**$_\mathbf{0}$, **S**$_\mathbf{1}$ |

TAB. 5.6 : Toponymes

Dans pareille situation, nous pouvons décider de nous fier aux indices de confiance. Si plusieurs étiquettes sémantiques en concurrence disposent d'un indice maximal, elles peuvent être listées dans le profil de lexie correspondant sous forme d'alternative, à l'aide du symbole |.

Parallèlement à cela, cette variation nous permet de constater que certaines dérivations lexicales sont indépendantes de l'organisation du lexique en espaces sémantiques projetée sur le lexique par l'ontologie du RL-fr. Les étiquettes ici observées y sont en effet éloignées l'une de l'autre. Tandis qu'entité sociale est l'instance d'une classe sémantique située directement sous la classe générale ENTITÉ, lieu géographique est l'instance d'une classe située à une profondeur de 3 de cette classe générale[10]. La décision de ne pas intégrer d'Attributs rendant compte de cet élément de description pour la mesure de similarité entre sommets lexicaux était donc judicieuse.

La répartition des FP, pour sa part, présente un cas d'accord parfait entre toutes les lexies. Une seule FP est utilisée ici, toujours couplée d'un indice de confiance de 100%. Elle peut donc être directement versée au profil de lexie en cours d'élaboration.

Pour finir, le détail des FL en jeu dans les liens sortants de chacune des lexies met en avant l'existance d'un triplet régulier : **Gener**, **A**$_\mathbf{0}$, **S**$_\mathbf{1}$. Les FL supplémentaires observées, propres à certaines lexies, sont exclues sans autre forme de procès. Contrairement aux étiquettes sémantiques et aux FP, les FL ne constituent pas des éléments de caractérisation lexicographique unique. Dans la grande majorité des cas, une lexie pour laquelle le travail lexicographique est avancé dispose de plusieurs liens de FL sortants. Dans le cas où les profils de lexies que nous cherchons à mettre au point seraient intégrés dans un éditeur lexicographique, nous pensons qu'il serait plus utile aux lexicographes de disposer d'un ensemble de suggestions de FL restreint

[10]L'étiquette lieu géographique est une instance de la classe LIEU_GÉOGRAPHIQUE, fille de la classe LIEU_PHYSIQUE, elle-même fille de la classe LIEU, située directement sous la classe générale ENTITÉ.

mais fiable plutôt que d'un ensemble vaste. En cela, nous estimons que les FL non partagées relèvent de particularités de chaque lexie et que nous ne pouvons juger de la pertinence de leur propagation a priori.

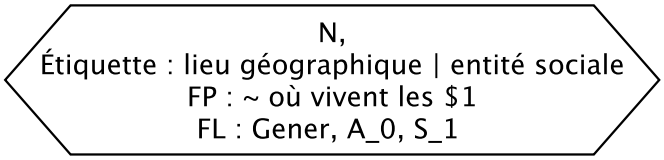



FIG. 5.14 : Profil de toponyme

L'ensemble de ces considérations nous amène à proposer le profil de lexie illustré par la figure 5.14. Les différents éléments de descriptions y sont présentés sous une forme adaptée à la lecture humaine. Ils devront cependant être enregistrés sous la forme des identifiants uniques référencés de la base de données du RL-fr, lors de l'implémentation de l'abstraction que nous esquissons ici[11].

Les éléments de descriptions disponibles pour les lexies nominales regroupées ici nous ont permis d'établir un profil complet. En interrogeant la base de données du RL-fr du 6 mai 2014 à l'aide d'une requête SQL, nous avons observé qu'aucune lexie supplémentaire ne correspondaient à ce profil. Réalisée sur la base de travail lexicographique du RL-fr le 17 octobre 2014, la même interrogation ne retourne plus que la seule lexie AFRIQUE[12]. Les liens de FL $\mathbf{S_1}$ des autres lexies ne sont en effet plus encodées dans leurs descriptions. Les liens de FL **Gener** ont pour leur part été supprimés de celles des locutions nominales ⌜AMÉRIQUE DU NORD⌝ et ⌜AMÉRIQUE DU SUD⌝. Par ailleurs, le vocable AMÉRIQUE DU NORD ne contient plus qu'une seule acception.

Ces observations nous mettent en garde au sujet de la validité de la configuration de dérivation lexicale que nous établissons ici. En effet, elle implique un lien de dérivation sémantique nominal de premier actant entre les toponymes et les gentilés[13] que les lexicographes ont jugé bon de supprimer. Les occurrences de ce lien que nous avions observées ont toutes été remplacées par un lien de dérivation prédicative sémantique nominal de premier actant, $\mathbf{S_1}$**Pred**, entre les adjectifs toponymiques et les gentilés.

### 5.6.2 Adjectifs toponymiques

Les descriptions des lexies de la seconde colonne disposent toutes de l'unique CG `adjectif`. La décision de lui substituer la méta-pdd utilisée lors du regroupement

[11] Aucune réflexion n'a pour l'instant été menée sur le format de représentation idéal des configurations de dérivations lexicales. Nous pensons cependant que l'ensemble de nos besoins sera compatible avec le standard RDF.

[12] Cette irrégularité de la lexie AFRIQUE par rapport aux autres typonymes a été signalée à l'équipe de lexicographes.

[13] Nous utilisons le terme de gentilés pour désigner les noms d'habitants d'un lieu.

des microstructures analogues convient donc parfaitement dans ce cas.

Les autres éléments de description lexicographique disponibles pour chacun de ces adjectifs sont présentés dans le tableau 5.7. Les exemples d'emploi qui les complètent sont au nombre de cinq. La description de la lexie AFRICAIN$_{Adj}$ **1** n'en contient en effet aucun.

| Lexies | Étiquettes sémantiques | FP | FL |
|---|---|---|---|
| AFRICAIN$_{Adj}$ | qui a une certaine caractéristique | [X] ∼ | **S$_0$** |
| AMÉRICAIN$_{Adj}$ **II.1** | | | **S$_0$** |
| SUD-AMÉRICAIN$_{Adj}$ | | [X] ∼ | **S$_0$** |
| ASIATIQUE$_{Adj}$ **1** | | | **S$_0$** |
| AUSTRALIEN$_{Adj}$ **II** | caractéristique sociale | [X] ∼ | **S$_0$** |
| EUROPÉEN$_{Adj}$ **I.1** | fait | [X] ∼ | **S$_0$** |

TAB. 5.7 : Adjectifs toponymiques

Comme nous pouvons le voir dans ce tableau, l'étiquetage sémantique proposé n'est pas satisfaisant. Trois lexies ne sont associées à aucune étiquette et les trois autres fournissent des informations contradictoires. De plus, les indices de confiance des associations en présence sont tous de 60%. Dans pareille situation, il semble préférable de n'en conserver aucune.

Le cas des FP est différent. Une complétion des descriptions lexicographiques qui n'en contiennent pas semble possible. En effet, les lexies AFRICAIN$_{Adj}$, SUD-AMÉRICAIN$_{Adj}$, AUSTRALIEN$_{Adj}$ **II** et EUROPÉEN$_{Adj}$ **I.1** sont toutes associées à la même instance de FP, avec un indice de confiance de 100%. Cette instance peut donc être intégrée au profil d'adjectif toponymique en cours d'élaboration.

Les liens de FL sortants présentent, pour leur part, un cas d'accord parfait. La présence d'une FL de nominalisation **S$_0$** peut donc être directement versée au profil de lexie établi.

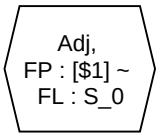



FIG. 5.15 : Profil d'adjectif toponymique

L'ensemble de ces considérations nous amène à proposer le profil d'adjectif toponymique illustré par la figure 5.15. Contrairement aux descriptions disponibles pour les toponymes, celles des adjectifs toponymiques ne nous ont pas permis d'établir un profil complet.

En interrogeant la base de données du RL-fr du 6 mai 2014, nous avons observé que 81 lexies supplémentaires correspondaient à ce profil. Huit d'entre elles étaient bien des adjectifs toponymiques, susceptibles de déclencher la configuration de dérivations lexicales que nous sommes en train d'élaborer : ALLEMAND$_{\text{Adj1}}$, ARCTIQUE$_{\text{Adj}}$, ANTARCTIQUE$_{\text{Adj}}$, ANTARCTIQUE$_{\text{Adj}}$, AUSTRALIEN$_{\text{AdjI.a}}$, BELGE$_{\text{Adj1}}$, EURASIEN, OCÉANIQUE, PARISIEN$_{\text{Adj}}$. Deux autres s'en rapprochaient : CONTINENTAL$_{\text{AdjI.1}}$, VILLAGEOIS$_{\text{Adj}}$. L'ensemble des 75 autres, en revanche, était très éloigné des adjectifs visés et très varié. Il comportait notamment les lexies adjectivales ANALOGIQUE **1** et PLUS BLANC QUE BLANC$_{\text{Adj}}$ **I**.

La même interrogation, réalisée sur la base de travail lexicographique du RL-fr le 17 octobre 2014, retourne des résultats comparables. Sur les 292 adjectifs correspondant au profil, 39 sont des adjectifs toponymiques, parmi lesquels il est intéressant de noter que les lexies AMÉRICAIN$_{\text{Adj}}$ **II.1** et ASIATIQUE$_{\text{Adj}}$ **1** sont désormais intégrés. Nous remarquons également la présence des quatre adjectifs de religion BOUDDHISTE$_{\text{Adj}}$ **1**, CHRÉTIEN$_{\text{Adj}}$ **1**, JUIF$_{\text{Adj}}$ **1** et MUSULMAN$_{\text{Adj}}$ **1**. La majorité des adjectifs retournés demeure toutefois inopportune.

Ces observations nous mettent en garde contre le déclenchement d'une configuration lexicale sur la base d'un seul profil de lexie, qui plus est incomplet. Il conviendrait sans doute de mettre au point une procédure permettant de distinguer, parmi les profils de lexies et les connexions lexicales de chaque configuration, certains éléments déclencheurs et d'autres, qui ne le seraient pas. Une autre solution consisterait à établir un seuil de nombre d'élements permettant ce déclenchement.

### 5.6.3 Gentilés

Les descriptions des lexies de la troisième colonne partagent toutes les CG `nom commun` et `masc`. AMÉRICAIN$_{\text{N}}$ **I.2**, SUD-AMÉRICAIN$_{\text{N}}$ et AUSTRALIEN$_{\text{N}}$ disposent chacun d'une caractéristique supplémentaire, qui semble avoir pour objectif de fournir la même information, mais sous une forme différente : `!s'écrit avec une majuscule initiale`, `s'écrit aussi « sans majuscules initiales »`, `s'écrit aussi sans majuscule initiale`. Cette caractéristique ne nous semble pas être particulière à ces trois lexies, mais relever plutôt d'une règle d'usage commune à tous les gentilés du français. Nous pourrions alors décider de l'intégrer au profil de lexie que nous souhaitons élaborer, mais il faudrait alors privilégier l'une des formes en présence. Nous ne disposons cependant d'aucun critère permettant d'implémenter un tel choix[14].

Nous resterons donc sur notre décision de substituer aux CG la méta-pdd utilisée lors du regroupement des microstructures analogues.

Les autres éléments de description lexicographique disponibles pour chacun des gentilés sont présentés dans le tableau 5.8. Les exemples d'emploi qui les complètent sont au nombre de sept. Cependant, la description de la lexie AMÉRICAIN$_{\text{N}}$ **I.2** n'en

[14]De plus, nous avons constaté que ces trois variantes d'une même information n'étaient plus présentes dans les descriptions des lexies AMÉRICAIN$_{\text{N}}$ **I.2**, SUD-AMÉRICAIN$_{\text{N}}$ et AUSTRALIEN$_{\text{N}}$ le 17 octobre 2014.

contient aucun.

| Lexies | Étiquettes sémantiques | FP | FL |
|---|---|---|---|
| AFRICAIN$_{N}$ | individu qui a un attribut [donné] | `X=1, qui est ART` ∼ | **Syn**$_{\supset^{sex}}$ |
| AMÉRICAIN$_{N}$ **I.2** | | | **Syn**$_{\supset^{sex}}$ |
| SUD-AMÉRICAIN$_{N}$ | individu qui a un attribut [donné] | `X=1, qui est ART` ∼ | **Syn**$_{\supset^{sex}}$ |
| ASIATIQUE$_{N,masc}$ | individu qui a un attribut [donné] | `X=1, qui est ART` ∼ | **Syn**$_{\supset^{sex}}$ |
| AUSTRALIEN$_{N}$ | individu qui a un attribut [donné] | `X=1, qui est ART` ∼ | **Syn**$_{\supset^{sex}}$ |
| EUROPÉEN$_{N}$ | individu qui a un attribut [donné] | `X=1, qui est ART` ∼ | **Syn**$_{\supset^{sex}}$ |

TAB. 5.8 : Gentilés

Comme nous pouvons le voir dans ce tableau, l'étiquetage sémantique et l'association lexie-FP sont tout aussi réguliers l'un que l'autre. Tandis que les lexies AFRICAIN$_{N}$, SUD-AMÉRICAIN$_{N}$, ASIATIQUE$_{N,masc}$, AUSTRALIEN$_{N}$ et EUROPÉEN$_{N}$ sont toutes associées à l'étiquette sémantique individu qui a un attribut [donné] et à la FP `X=1, qui est ART` ∼, la lexie AMÉRICAIN$_{N}$ **I.2** ne dispose d'aucun de ces deux types d'éléments. Nous pouvons donc considérer qu'il s'agit là d'un cas de complétion ne présentant aucune difficulté.

Concernant les liens de FL sortants, nous nous trouvons face à un cas simple d'accord parfait. Nous pouvons donc intégrer au profil de lexie en cours d'élaboration la présence obligatoire d'un lien de synonymie plus riche relative au sexe, **Syn**$_{\supset^{sex}}$.

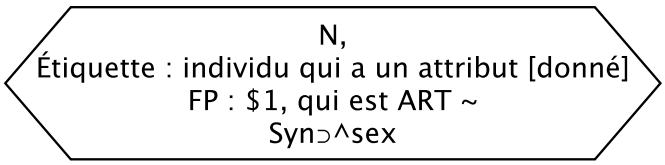



FIG. 5.16 : Profil de gentilé

L'ensemble de ces considérations nous amène à proposer le profil illustré par la figure 5.16. Tout comme pour les toponymes, les descriptions disponibles pour les gentilés nous ont permis d'aboutir à la création d'un profil complet. Il comporte cependant un seul lien de FL sortant. Nous ne sommes pas en mesure de déterminer automatiquement si cette faible connectivité lexicale est une particularité des gentilés ou si cela est dû à la faible avancée du travail lexicographique qui leur a été consacré. Cependant, d'après leur statut, nous savons que la plupart de ces lexies n'ont fait l'objet d'aucun travail lexicographique particulier et que seule la lexie AFRICAIN$_{N}$ était en cours de traitement le 6 mai 2014.

En interrogeant la base de données du RL-fr de cette même date, nous avons observé que sept lexies supplémentaires correspondaient à ce profil. Une seule d'entre elles étaient un gentilé : ALLEMAND$_{N}$. Il est d'ailleurs intéressant de constater que

cette lexie peut être rapprochée de la lexie ALLEMAND$_{\text{Adj}}$ **1** que nous avons observée précédement. Deux autres résultats semblaient intéressants, il s'agissait des lexies ABORIGÈNE$_{\text{N,masc}}$ **b** et CONTINENTAL$_{\text{N}}$. Les quatre autres étaient sans aucun rapport avec les entités géographiques : BACHELIER$_{\text{N}}$, ⌜BON À RIEN$_{\text{N}}$⌝, ⌜CHEF DE BORD$_{\text{N,masc}}$⌝ et DORMEUR **b**.

La même interrogation, réalisée sur la base de travail lexicographique du RL-fr le 17 octobre 2014, ne retourne plus que cinq résultats : ABORIGÈNEN,MASC **b**, BACHELIER$_{\text{N}}$, ⌜BON À RIEN$_{\text{N}}$⌝, AFRICAIN$_{\text{N}}$, CONTINENTAL$_{\text{N}}$. Les gentilés ALLEMAND$_{\text{N}}$, AUSTRALIEN$_{\text{N}}$, ASIATIQUE$_{\text{N,masc}}$, EUROPÉEN$_{\text{N}}$ et SUD-AMÉRICAIN$_{\text{N}}$ sont désormais associés à l'étiquette sémantique individu[15].

Tout comme nous l'avons vu pour les toponymes, ces observations nous mettent en garde à propos de la validité de la configuration de dérivation lexicale que nous établissons ici. Bien que les connexions en jeu ne soient, dans ce cas, pas remises en cause, nous constatons que la prise en compte de lexies non travaillées et en cours de traitement lexicographique pour établir des configurations de dérivations lexicales peut s'avérer problématique. Si leur exploitation est envisagée pour enrichir semi-automatiquement le RL-fr, elle pourrait amener à la propagation d'erreurs. En revanche, leur exploitation pour vérifier la cohérence du réseau demeure intéressante.

### 5.6.4 Configuration de dérivations lexicales toponymiques

Nous venons d'établir un profil pour chacune des positions des sous-graphes constitués de triplets (toponyme, adjectif toponymique, gentilé) que nous avions préalablement regroupés comme relevant de microsctructures lexicales analogues. Nous sommes donc à présent en mesure de proposer un résultat complet d'abstraction de configuration de dérivations lexicales en intégrant ces profils au squelette de connexions lexicales de n'importe lequel de ces sous-graphes. La figure 5.17 propose une représentation du résultat de cette dernière étape.

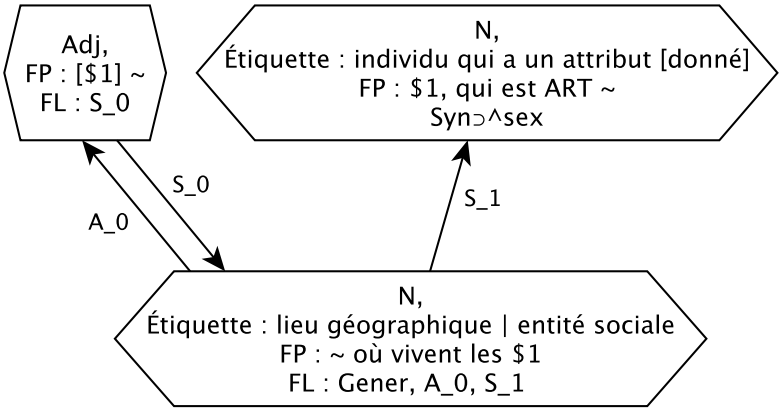



FIG. 5.17 : Configuration de dérivations lexicales toponymiques

[15]La lexie AMÉRICAIN$_{\text{N}}$ **I.2** disposait également de l'étiquette individu en date du 17 octobre 2014.

Cette représentation n'inclut pas de distinction d'éléments déclencheurs que nous avons évoqué précédemment. Il nous semble que le toponyme en présence est le plus à même de jouer un tel rôle et qu'il serait intéressant que l'encodage d'informations correspondant à son profil permette aux lexicographes d'accéder à un ensemble de suggestions de configurations dans lesquelles il pourrait trouver sa place. Cependant, nous n'avons pas établi de règles implémentables permettant de distinguer automatiquement les déclencheurs ; noyer les lexicographes sous les informations en considérant chaque élément comme étant déclencheur ne nous paraît pas judicieux. De la même façon, nous pourrions envisager l'intégration semi-automatique de liens à partir d'une telle lexie, mais il faudrait pour cela être en mesure d'émettre des hypothèses sur les lexies qu'elle devrait sélectionner comme cibles. De telles exploitations ne sont donc pas envisageables pour l'instant. D'autres applications nous semblent en revanche plus directement accessibles.

En premier lieu, nous pouvons envisager de rechercher les triplets de lexies correspondant partiellement à une telle configuration et d'en compléter les descriptions. Pour cette configuration précise, nous avons vu qu'aucune projection de ce type n'est envisageable sur la version du RL-fr du 17 octobre 2014, où les descriptions des lexies ayant permis sa mise au point ont été modifiées. En revanche, son exploitation n'est pas sans intérêt dans le cadre de la version dont elle a été extraite. Ainsi, nous pourrions ajouter la FP `[X]` $\sim$ aux descriptions des lexies AMÉRICAIN$_{\text{Adj}}$ **II.1** et ASIATIQUE$_{\text{Adj}}$ **1**, ainsi que l'étiquette sémantique individu qui a un attribut [donné] et la FP `X=1, qui est ART` $\sim$ à celle de la lexie AMÉRICAIN$_{\text{N}}$ **I.2**. De tels ajouts seraient associés à un indice de confiance de 50%, pour signaler aux lexicographes qu'il s'agit d'informations nécessitant une validation manuelle.

Nous pourrions également envisager de compléter le couple (ALLEMAND$_{\text{Adj}}$ **1**, ALLEMAND$_{\text{N}}$) mis en avant dans les sections 5.6.3 et 5.6.2. Le lien de nominalisation, $\mathbf{S_0}$, encodé dans la descriptions de l'adjectif toponymique de ce couple a d'ores et déjà pour cible le toponyme ALLEMAGNE qui manque pour obtenir le triplet qui nous intéresse. De plus, la description de cette lexie contient la CG `nom propre` et un lien d'adjectivation inverse, $\mathbf{A_0}$ qui appartient également à la configuration que nous avons établie. Le reste de sa description est vide. Elle ne contient ni étiquette sémantique, ni FP, ni exemple lexicographique. Elle est donc un bon candidat à la complétion. Il n'existe cependant aucun lien de FL encodant une relation entre la lexie ALLEMAND$_{\text{N}}$ et ALLEMAND$_{\text{Adj}}$ **1** ou ALLEMAGNE. Nous nous trouverions donc dans une situation d'intégration semi-automatique de liens, pour laquelle nous manquons de critères de sélection des cibles. Dans le cas particulier évoqué ici, nous avons constaté que les descriptions des lexies ALLEMAND$_{\text{Adj}}$ **1** et ALLEMAND$_{\text{N}}$ disposaient toutes les deux d'informations morphologiques et que les bases morphologiques qui leur étaient associées[16] étaient identiques. Nous aurions donc pu envisager de nous appuyer sur ces informations pour parvenir à nos fins. Cependant, il s'agit là d'un cas très particulier. Nous ne saurions donc nous satisfaire d'une telle solution.

Nous avons alors souhaité tester s'il était possible de nous appuyer sur les classes de proportions analogiques équivalentes énonçables à partir des sous-graphes ayant permis l'abstraction de la configuration courante pour sélectionner ALLEMAND$_{\text{N}}$

[16] Un exemple d'une telle base est visible dans la figure 2.11, section 2.2.4.3 du chapitre 2.

comme cible d'un lien de dérivation sémantique de premier actant, $\mathbf{S_1}$, ayant pour source ALLEMAGNE. Rappelons-le, à partir de chaque couple de sous-graphes analogues, il est possible d'énoncer un ensemble de classes de proportions analogiques. Ainsi, à partir des triplets (⌜AMÉRIQUE DU NORD **II**⌝, AMÉRICAIN$_{\text{Adj}}$ **II.1**, AMÉRICAIN$_{\text{N}}$ **I.2**) et (AUSTRALIE **II**, AUSTRALIEN$_{\text{Adj}}$ **II**, AUSTRALIEN$_{\text{N}}$), nous pouvons énoncer les trois classes suivantes :

⌜AMÉRIQUE DU NORD **II**⌝ : AMÉRICAIN$_{\text{Adj}}$ **II.1** :: AUSTRALIE **II** : AUSTRALIEN$_{\text{Adj}}$ **II**,
⌜AMÉRIQUE DU NORD **II**⌝ : AMÉRICAIN$_{\text{N}}$ **I.2** :: AUSTRALIE **II** : AUSTRALIEN$_{\text{N}}$,
AMÉRICAIN$_{\text{Adj}}$ **II.1** : AMÉRICAIN$_{\text{N}}$ **I.2** :: AUSTRALIEN$_{\text{Adj}}$ **II** : AUSTRALIEN$_{\text{N}}$.

Nous avons émis l'hypothèse, dans la section 2.2.3.1 du chapitre 2, qu'un certain nombre de rapports mis en évidence par ces proportions pouvaient trouver une expression dans une dimension topologique. Pour la classe ⌜AMÉRIQUE DU NORD **II**⌝ : AMÉRICAIN$_{\text{N}}$ **I.2** :: AUSTRALIE **II** : AUSTRALIEN$_{\text{N}}$, nous supposions ainsi que le rapport mis en évidence par l'analogie « ⌜AMÉRIQUE DU NORD **II**⌝ **est à** AUSTRALIE **II** **ce que** AMÉRICAIN$_{\text{N}}$ **I.2 est à** AUSTRALIEN$_{\text{N}}$ » pouvait être exprimé en terme de position des sommets dans le graphe lexical. De plus, si ALLEMAND$_{\text{N}}$ était bien la cible souhaitée pour le lien de dérivation sémantique du premier actant à injecter, nous estimions que ce rapport devait être conforme à ceux mis en évidence par les proportions ⌜AMÉRIQUE DU NORD **II**⌝ : ALLEMAGNE :: AMÉRICAIN$_{\text{N}}$ **I.2** : ALLEMAND$_{\text{N}}$ et ALLEMAGNE : AUSTRALIE **II** :: ALLEMAND$_{\text{N}}$ : AUSTRALIEN$_{\text{N}}$.

Nous avons alors choisi de nous intéresser aux plus courtes chaînes[17] existant dans le RL-fr entre l'ensemble des toponymes AFRIQUE, ⌜AMÉRIQUE DU NORD **II**⌝, ⌜AMÉRIQUE DU SUD⌝, ASIE, AUSTRALIE **II** et EUROPE **I**, ainsi qu'entre l'ensemble des adjectifs toponymiques et l'ensemble des gentilés qui leur étaient associés. Nos premières observations étaient encourageantes. La longueur de ces plus courtes chaînes était régulière. Les distances entre les adjectifs toponymiques d'une part et les gentilés d'autre part étaient identiques, celles entre les toponymes étaient légérement inférieures, car ils partageaient des hyperonymes communs : toponymes séparés par deux arcs, adjectifs toponymiques séparés par quatre arcs, gentilés séparés par quatre arcs. Néanmoins, nous n'avons pas été en mesure de poursuivre dans cette voie, car les lexies ALLEMAGNE, ALLEMAND$_{\text{N}}$ et ALLEMAND$_{\text{Adj}}$ **1** étaient situées dans une composante connexe[18] distincte de celle qui contenait les sous-graphes analogues. Il n'existait donc aucune chaîne entre les lexies ⌜AMÉRIQUE DU NORD **II**⌝ et ALLEMAGNE, pas plus qu'entre les lexies AMÉRICAIN$_{\text{N}}$ **I.2** et ALLEMAND$_{\text{N}}$. La question de l'exploitation des classes de proportions analogiques et des rapports entre sommets lexicaux établis selon une dimension topologique pour la complétion de sous-graphes et l'injection semi-automatique de relations reste donc en suspens.

En second lieu, il nous semble que de telles configurations pourraient être exploitées dans un cadre pédagogique. Nous avons signalé dans le chapitre 1, section 1.3.1, que les sciences de l'éducation s'intéressent au raisonnement analogique

[17]Une plus courte chaîne correspond à la plus petite suite consécutive d'arcs par lesquels il est nécessaire de passer pour se rendre d'un sommet à un autre, sans prendre en compte l'orientation des arcs. Il est possible que plusieurs chaînes minimales existent entre deux sommets.

[18]La notion de composantes connexes est présentée dans la section 2.3.1 du chapitre 2.

en tant qu'outil pédagogique. Nous avons alors cité les travaux de Wong (1993), qui développe l'idée selon laquelle la construction d'analogies personnelles permet à un apprenant de s'approprier de nouveaux concepts scientifiques en les situant par rapport à ses propres connaissances et qu'elle favorise l'abstraction.

Nous avons eu l'occasion, dans le cadre d'un atelier de vulgarisation scientifique organisé avec des enfants de classes de CE2 et CM1, de tester la co-construction d'une abstraction à partir de l'observation de la connectivité d'un ensemble de lexies dénotant des sports. Cette abstraction a ensuite été projetée par les enfants sur une nouvelle situation. Menée de manière informelle, cette expérience a bien fonctionné. En construisant les analogies préalables à l'abstraction, les enfants se sont appropriés de nouveaux concepts lexicaux, qu'ils ont su exploiter par la suite.

# Conclusion

Nous avons présenté dans ce chapitre une expérience complète d'exploration du RL-fr par raisonnement analogique sur les motifs locaux de taille 3. Nous avons ainsi établi l'existence de microstructures analogues relevant de mêmes configurations de dérivations lexicales.

La première observation que nous pouvons tirer de cette expérience est qu'une très faible quantité des occurrences de motifs collectées s'est avérée analogues. Parmi les 162 motifs initialement repérés comme disposant de plus de deux occurrences, seuls dix étaient en œuvre dans les sous-graphes conservés. Ils étaient cependant tous parmi les vingt-cinq plus fréquents. Cette particularité pourrait être exploitée par la suite pour réduire le champ d'investigation d'une telle exploration. Il est cependant possible qu'elle soit liée à la faible avancée des descriptions lexicographiques encapsulées dans les sommets lexicaux du RL-fr. Pour compléter cette analyse, d'autres expériences devront être menées, sur des graphes lexicaux plus aboutis tels que les réseaux JeuxDeMots et WordNet que nous avons évoqués dans la section 2.1.1 du chapitre 2.

Comme nous l'espérions, l'abandon des composantes connexes au profit des motifs locaux nous a permis de valider les hypothèses qui étaient préalablement restées en suspens. Chacun des Attributs que nous avons exploités pour comparer les descriptions lexicographiques encapsulées dans les sous-graphes partageant la même structure de connexions lexicales s'est avéré pertinent. Nous avons cependant entrevu que le calcul du rapport arithmétique entre le nombre de liens entrants et sortants de chaque sommet lexical pourrait être amélioré en prenant en considération la présence de mentions spécifiant les marques d'usage des cibles de certains de ces liens. Nous pourrions ainsi exclure de ce calcul les liens en direction de lexies rares ou vieillis.

Ce changement de délimitation de sous-graphes nous a amenée à l'ajout d'un nouveau critère au cours de l'étape de comparaison des sommets lexicaux. Nous avons établi que pour relever d'analogies pertinentes, les sous-graphes comparés devaient non seulement partager un minimum de similarités d'Attributs complètes égal au nombre de sommets qui les composent, mais que ces similarités devaient

mettre en jeu un minimum de lexies distinctes. Nous avons pu observer que ce nouveau critère engendrait des scissions non souhaitables et qu'il nécessitait l'ajout d'un post-traitement. Cependant, nous avons souhaité le conserver, car il limitait les cas de regroupements analogiques peu pertinents. Nous avons mené, en parallèle, une expérience dépourvue de ce nouveau critère. À partir des motifs locaux de taille 4 de la plus grande composante fortement connexe du RL-fr sans lien d'inclusion formelle du 6 mai 2014, nous avons collecté 485 944 occurrences de motifs, pour 1 610 motifs distincts. L'ensemble de ces occurrences ne mettait en jeu que 2 969 lexies. L'observation d'un échantillon de 29 818 groupes de microstructures analogues obtenus au cours de cette expérience — cet échantillon représentait près de 50% des groupes de sous-graphes partageant les mêmes connexions lexicales — nous a montré que plus de 87% d'entre eux contenaient des sous-graphes ayant tous en commun trois lexies sur quatre.

Nous avons évoqué, dans ce chapitre, un certain nombre de problèmes liés à la prise en compte des lexies peu décrites. Nous nous sommes alors intéressée à une version allégée du RL-fr, ne comportant que les 264 lexies en cours de vérification ou considérées comme entièrement décrites de la version du 6 mai 2014. Cet ensemble aurait pu constituer une base d'apprentissage de configurations de dérivations lexicales que nous aurions ensuite pu projeter sur l'ensemble du réseau. Les performances matérielles de notre procédure en auraient été améliorées et les profils de lexies intégrés aux configurations auraient sans aucun doute été plus aboutis. Cependant, cet échantillon ne se prêtait pas à l'exploration des microstrucures du réseau. En effet, si les lexies qu'ils comportaient étaient effectivement entièrement décrites, les relations de copolysémie qu'elles entretiennent n'avaient alors pas encore été toutes encodées. De plus, les liens de FL qui les impliquaient avaient majoritairement pour sources ou pour cibles des lexies absentes de l'échantillon. Ainsi, en nous intéressant aux motifs de taille 3, nous n'avons pu collecter aucun groupe de microstructures analogues contenant des sous-graphes composés exclusivement de lexies distinctes. En relâchant la contrainte de nombre de lexies distinctes, nous obtenions trente deux groupes de sous-graphes analogues, qui partageaient tous deux lexies. De plus, près de la moitié d'entre eux impliquaient la lexie FAIRE **II.1** [*Il fait du ping-pong*].

Aucun de ces groupes n'impliquait de lexies appartenant aux vocables du champs lexical du sport auxquels nous nous étions intéressée en conclusion du chapitre 3. Les vocables $\text{BASKET}_{\text{N,masc}}$, FOOTBALL, PING-PONG, RUGBY, $\text{TENNIS}_{\text{N,masc}}$ et VOLLEY-BALL étaient pourtant bien entièrement intégrés à cet échantillon. Les mesures de similarités réalisées entre les descriptions de leurs acceptions[19] avaient toutes abouties à un score situé entre 0,76 et 0,95. Cette observation nous a amené à penser que la mise en place d'une étude des structures polysémiques analogues nécessiterait une modification de l'ensemble d'Attributs exploité pour les configurations de dérivations lexicales impliquant également des liens de FL. Nous pourrions ainsi envisager plusieurs possibilités. L'une d'entre elles, suggérée par les patrons proposés par Barque (2008) pour les domaines de polysémie des sentiments et des animaux, serait de se concentrer sur les éléments de description relatifs aux définitions des

[19]Nous parlons ici de leurs acceptions analogues — au sens commun du terme — deux à deux, telles que $\text{BASKET}_{\text{N,masc}}$ **2** [*Veux-tu faire un basket demain soir ?*] et FOOTBALL **3** [*Après l'école, nous avons fait un football avec les copains.*].

lexies. Une seconde, suggérée par les travaux de Sikora (2014) sur la dérivation sémantique régulière des verbes entre un sens d'achèvement et un sens statif, serait de sélectionner un sous-ensemble de FL pertinentes pour la définition des Attributs, couplé à l'exploitation des étiquettes sémantiques et des FP.

La réflexion que nous avons débutée sur la question de l'abstraction automatique pourrait être exploitée aussi bien pour ce nouveau type de configurations que pour celles impliquant des liens de FL. Les applications d'instrumentation du travail lexicographique et d'activités pédagogiques que nous avons envisagées pourrait également les inclure.

Les exemples lexicographiques ont été laissés de côté lors de l'abstraction de configurations de dérivations lexicales. Nous avons cependant envisagé de mesurer la spécificité de l'ensemble des exemples encapsulés dans les sommets occupant une même position par rapport au reste de la collection d'exemples du RL-fr. Cette mesure aurait pu être effectuée à l'aide de l'outil de textométrie TXM (Heiden et al., 2010). Faute de temps, la mise en place de la procédure nécessaire à ce travail n'a malheureusement pas été réalisée. La pertinence de l'ajout d'une telle information et les questions relatives à son intégration aux configurations n'ont donc pas pu être abordées.

# Conclusion

Nous avons proposé, dans cette thèse, une méthode d'exploration de graphes lexicaux par raisonnement analogique. Nous avons défini cette exploration comme un regroupement de structures unifiables. Les sommets du graphe lexical parcouru s'apparentent alors à des objets disposant d'un certain nombre d'Attributs, disponible dans leur description lexicographique. Ils entretiennent des Relations, représentées par les arcs.

Nous avons vu qu'une étude préalable du réseau était nécessaire à cette exploration, afin de définir avec précision la nature des objets qu'il met en relation et de définir un ensemble d'Attributs permettant leur comparaison. Nous avons évoqué à plusieurs reprises que le choix de ces Attributs détermine la dimension selon laquelle les regroupement sont effectués. Ainsi, l'inclusion d'étiquettes sémantiques engendrerait la prise en compte de l'organisation du lexique selon les champs sémantiques défnis dans la ressource. Nous nous sommes concentrée ici sur l'identification de structures de dérivations lexicales réccurrentes. Les Attributs sélectionnés étaient alors tous en lien avec la connectivité des sommets :

- un Attribut rendant compte de leur partie du discours, dont nous savons qu'elle détermine en partie la dérivation syntaxique des lexies ;
- des Attributs rendant compte des familles de FL en jeu dans leurs liens entrants et sortants, autrement dit les grands types de relations syntagmatiques et paradigmatiques qu'ils entretiennent avec le reste du lexique ;
- un Attribut rendant compte du rapport arithmétique entre le nombre de liens entrants et le nombre de liens sortants, fournissant un indice sur leur rôle dans l'organisation topologique du réseau.

Bien que nous nous soyons limitée à l'exploration du RL-fr, nous pensons que la méthode développée ici peut être adaptée à d'autres réseaux lexicaux. Nous envisageons ainsi, en collaboration avec Yann Desalle, membre du groupe informel de chercheurs en informatique, linguistique et sciences cognitives Proxteam[20] actuellement en contrat post-doctoral à l'ATILF, une exploration du réseau JeuDeMots (Lafourcade et Joubert, 2008, 2010). Une réflexion sur les éléments constitutifs de ce réseau devra alors être menée. Certaines des relations qu'il met en œuvre, telles que les lieux typiques des actions ou les acteurs des processus et des événements, sont proches de certaines fonctions lexicales Sens-Texte encodées dans le RL-fr. D'autres, telles que la couleur d'un objet ou les personnages d'une œuvre, sont d'une toute autre nature. Cette distinction aura nécessairement des conséquences sur la nature

[20] `http ://www.irit.fr/Proxteam`

des regroupements analogiques effectués. De plus, contraitement au RL-fr, il s'agit d'un graphe pondéré. Il faudra alors déterminer de quelle manière cette pondération doit être prise en compte. Une adaptation à l'exploration du WordNet de Princeton (Fellbaum, 1998) soulèverait également un certain nombre de questions. La nature de ses sommets, qui sont des ensembles de synonymes, son découpage en parties du discours et le caractère majoritairement hiérarchique des relations qu'il encode auraient eux aussi une incidence sur les regroupements effectués. La variation des éléments de descriptions disponibles en fonction des parties du discours, telles que les patrons syntaxiques des ensembles de synonymes verbaux ou les traits syntaxiques des adjectifs, devraient également être considérés.

La nature des relations encodées dans le RL-fr nous a, pour sa part, permis d'envisager l'émergence de modèles de dérivations lexicales régulières. Nous avons ainsi défini l'objectif de la majeure partie de nos explorations. Il s'agissait de regrouper des sous-graphes extraits de la ressource, qui mettaient en jeu des relations de dérivations syntaxiques et sémantiques communes et dont les lexies partageaient deux à deux un ensemble d'Attributs identiques.

Nous avons mis en avant que les strictes similarités de connexions lexicales et d'Attributs en jeu dans les sous-graphes considérés comme analogues permettaient, presque systématiquement, de s'assurer que les couples de sommets en similarité d'Attributs complète occupaient la même place dans les microstructures regroupées. Elles garantissaient ainsi la possibilité d'énoncer un ensemble de proportions analogiques entre leurs sommets lexicaux.

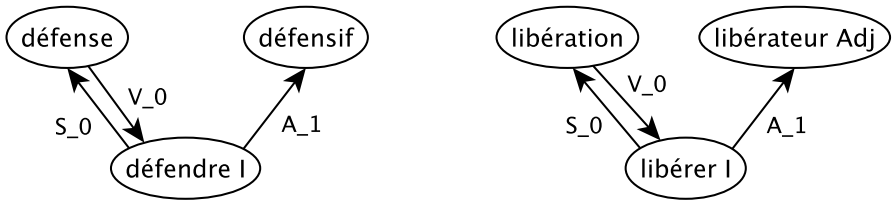



FIG. 5.18 : Sous-graphes analogues

Ainsi, le regroupement des deux sous-graphes illustré par la figure 5.18, obtenu lors de l'expérience détaillée au chapitre 5, permet d'énoncer trois classes d'analogies équivalentes :

LIBÉRATION : LIBÉRER I :: DÉFENSE : DÉFENDRE I,
LIBÉRATEUR$_{\text{Adj}}$ : LIBÉRER I :: DÉFENSIF : DÉFENDRE I,
LIBÉRATION : LIBÉRATEUR$_{\text{Adj}}$ :: DÉFENSE : DÉFENSIF.

Nous avons constaté que l'énonciation des rapports en jeu dans une parties des analogies équivalentes d'une même classe pose cependant problème. Ainsi, nous pouvons difficilement énoncer ce que « LIBÉRATION est à DÉFENSE », pas plus que ce que « LIBÉRER I est à DÉFENDRE I ». Une telle conformité de rapport ne semble par ailleurs pas pertinente pour les lexicographes. Nous observons également ici que la

troisième classe d'analogie s'établit selon un rapport qui n'est pas encodée dans le réseau. Ce que « LIBÉRATION est à LIBÉRATEUR$_{Adj}$ » et « DÉFENSIF à DÉFENSE » n'est pas interprétable automatiquement à partir des arcs en présence. Une telle conformité de rapport peut cependant intéresser les lexicographes. Elle peut ainsi les amener à compléter la description des lexies LIBÉRATEUR$_{Adj}$ et DÉFENSE en y intégrant des liens de dérivation sémantique adjectivale de premier actant $\mathbf{A_1}$ ayant pour cibles, respectivement, LIBÉRATEUR$_{Adj}$ et DÉFENSIF.

Une fois de tels regroupement effectués, nous avons pu envisager l'abstraction de modèles. Nous avons baptisé ces modèles des *configurations de dérivations lexicales* et les avons définis comme des ensembles de relations orientées entre profils de lexies détaillant les caractéristiques nécessaires à leur déclenchement. Nous avons constaté que l'abstraction de telles configurations sur un réseau lexical en cours de développement posait différents problèmes. D'une part, les descriptions lexicographiques disponibles peuvent être incomplètes ou contradictoires. D'autre part, certaines informations encodées peuvent être erronées. Nous avons alors proposé des solutions simples pour rémédier à la première de ces difficultés. Elles permettent par ailleurs tout à la fois de prendre en compte les particularités de chaque lexie et d'envisager une complétion semi-automatique de certaines descriptions. La seconde difficulté pose le problème de la validitié des configurations élaborées. Elle met cependant en avant une autre application intéressante pour les lexicographes. En fournissant un ensemble systématique de microstructures erronées, notre approche leur permet de corriger conjointement l'ensemble des erreurs de façon cohérente. À l'inverse, en fournissant des configurations de dérivations lexicales approuvées par les lexicographes, elle pourrait constituer une étape préliminaire à l'injection semi-automatique de liens de dérivations dans la ressource. Cette application nécessiterait cependant que nous soyons en mesure d'émettre des hypothèses fiables sur les sommets que devraient sélectionner pour cibles les lexies disposant des caractéristiques suffisantes au déclenchement d'une configuration. Une collaboration est également envisagée avec Yann Desalle sur ces aspects.

Sur une ressource complète, ou sur une zone « sûre » délimitée d'une ressource en cours de développement, de telles abstractions permettraient d'envisager une instrumentation du travail lexicologique. Nous avons ainsi évoqués les travaux de Sikora (2014) sur la régularité du phénomène de lexicalisation, par certains verbes, d'un sens transitionnel de type achèvement et d'un sens statif. Cette étude des mécanismes sémantiques en œuvre dans une dérivation polysémique régulière pourrait bénéficier d'une exploration par raisonnement analogique particulière. Il s'agirait de déterminer avec précisions les Attributs pertinents à considérer. En nous appuyant sur ses premières observations, nous pourrions tester la pertinence d'Attributs tels que les étiquettes sémantiques, les FP et la sous-sélection des liens de FL paradigmatiques. Seules les dérivations de co-polysémie pouraient être pris en compte pour la comparaison des Relations et l'élaboration des modèles par abstraction. Il s'agirait alors de configuration de dérivations sémantiques de copolysémie.

Nous avons également évoqué les travaux de Barque (2008) sur la mise au point de patrons pour les domaines de polysémie des sentiments et des animaux. Les patrons ainsi élaborés suggèrent qu'une exploration particulière du RL-fr, mettant en

jeu un ensemble d'Attributs composé d'éléments relatifs aux définitions des lexies permettrait de dégager de nouveaux patrons. Lors des expériences présentées dans cette thèse, les paraphrases définitionnelles n'étaient pas encore disponibles dans les descriptions des lexies. Leur implémentation dans l'éditeur lexicographique couplé au RL-fr est cependant actuellement en voie d'achèvement. Une fois que de telles paraphrases seront disponibles dans un nombre significatif de descriptions partageant le même domaine sémantique, il deviendra envisageable d'effectuer un apprentissage automatique par analogie des patrons polysémiques de ce domaine. Nous pourrons alors projeter les connaissances acquises au cours de cet apprentissage sur l'ensemble du réseau pour y suggérer la création d'acceptions et la complétion de descriptions lexicographiques.

En dernier lieu, nous souhaitons évoquer la possibilité d'exploiter la méthode de raisonnement analogique mise au point dans cette thèse dans le cadre de l'étude de la qualification des néologismes à laquelle nous avons précédemment participé (Ollinger et Valette, 2008 ; Reutenauer et al., 2011). Lors de la réalisation des travaux consacrés à l'observation de la néologie de sens, présentés dans Reutenauer et al. (2011), nous explorions la combinatoire de noms et d'adjectifs que nous supposions employés dans un sens nouveau. Cette étude était réalisée sur un corpus thématique d'articles de presse quotidienne portant sur la crise financière et couvrant la période de septembre 2008 à février 2009. Nous avons ainsi observé l'adjectif *toxique* dans son nouvel usage exprimant la désapprobation du locuteur dans des collocations telles que *crédit toxique* ou *emprunt toxique*. À partir des résultats d'une analyse des spécificités du corpus par le logiciel TermoStat (Drouin et al., 2006), nous avons sélectionné ses cooccurrents nominaux les plus spécifiques. Nous avons donc retrouvé *crédit* et *emprunt*, auquels sont venus s'ajouter *produit* et *titre*. Une recherche des cooccurrents de ces derniers a montré que le plus spécifique d'entre eux été l'adjectif *financier*. Ce qui nous a frappé à l'époque de cette étude, et qu'il nous semble encore intéressant de relever aujourd'hui, c'est que cet adjectif *financier* entretient une relation privilégiée avec le nom *organisme*. Or, la structure polysémique du vocable ORGANISME comporte à la fois une acception biologique, comme l'acception prélablement lexicalisée de *toxique* et une acception relative aux institutions, notamment financières. Nous pensons que cette particularité n'est pas le fruit du hasard. Au contrainte, nous émettons l'hypothèse que la dérivation sémantique de copolysémie activée à l'occasion de la crise financière était déjà potentiellement présente dans le lexique du français. Il nous semble qu'une exploration de ressources lexicographiques par raisonnement analogique dans le but de définir des patrons polysémiques permettrait de rendre compte telles potentialités lexicales. Les néologismes de sens détectés en corpus pourraient alors être intégrés à ces ressources en bénéficiant de l'ensemble des informations lexicographiques détaillées dans le profil de lexie dont ils prendraient la place dans le patron polysémique. Ainsi, tout comme l'analogie formelle est à l'œuvre dans la création lexicale de nouveaux signifiants selon Saussure (1916), l'analogie sémantique serait à l'œuvre dans la création de néologismes de sens.

D'une façon comparable à celle dont Boussidan et al. (2009) étudient l'évolution dans le temps des réseaux sémantiques des néologismes en corpus, entre leur apparition et la stabilisation de leur lexicalisation, l'étude de l'évolution des réseaux

sémantiques lors de l'intégration de néologismes aux ressources lexicographiques permettrait de tester cette hypothèse. Il s'agirait alors de mesurer l'impact de cette intégration sur la topologie du réseau. À l'échelle de sa microstrucutre, nous pourrions notamment nous intéresser aux configurations de dérivations impliquées dans les connexions que la nouvelle lexie entretient avec ses proches voisins. À une échelle plus macroscopique, nous pourrions estimer l'évolution des espaces sémantiques que cette intégration engendre, à l'aide de la méthode de mesure de proximité sémantique développée par Gaume (2004). Une telle étude suposerait cependant d'accéder à une ressource lexicale du français formalisée selon les modèles de la lexicographie contemporaine et ayant déjà bénéficiée, a minima, d'une vague d'intégration de néologismes. À notre connaissance, une telle ressource n'est, pour l'heure, pas accessible.

# Annexe

La liste de fonctions lexicales proposée dans le document qui va suivre n'a pas pour ambition d'être exhaustive. Elle a été conçue comme un guide, permettant de se familiariser avec l'interprétation des FL et de faciliter la lecture des résultats des différentes expériences présentées au cours de cette thèse. Elle a été rédigée par Alain Polguère, que nous remercions cordialement pour ce travail.

# Liste de fonctions lexicales fréquemment utilisées dans le Réseau Lexical du Français (RL-fr)

Alain Polguère
alain.polguere@univ-lorraine.fr

27 octobre 2014

## Table des matières



## 1 Introduction

- Cette liste **non exhaustive** est inspirée de la liste de fonctions lexicales standard simples présentée dans (Mel'čuk et coll., 1995, pages 129–148).
- Dans ce qui suit, $x$ renvoie à la lexie argument de la fonction lexicale et $y$ à une valeur de l'application de la fonction lexicale à la lexie $x$. Par exemple, dans

  $$\textbf{Syn}(\textit{voiture}) = \textit{automobile}$$

  *voiture* est $x$ et *automobile* est $y$.
- Le sigle *PDD* est employé pour *partie du discours*.
- La liste des valeurs $y$ obtenues pour chaque application d'une fonction lexicale n'est pas exhaustive. Dans la plupart des cas, un seul $y$ est donné.

## 2 Fonctions lexicales paradigmatiques

### 2.1 Relations fondamentales

1. **Syn** : synonyme

| **PDD** $x$ | V | N | Adj | Adv |
|---|---|---|---|---|
| **PDD** $y$ | V | N | Adj | Adv |

Lexie fonctionnant comme paraphrase de $x$
**Syn**(*voiture*) = *automobile*
La synonymie, comme toute relation de dérivation sémantique, peut être approximative. On distingue trois types standard de synonymes approximatifs :

(a) synonymes moins spécifiques (sens de $y$ est inclus dans le sens de $x$) – $\textbf{Syn}_{\subset}$
$\textbf{Syn}_{\subset}(\textit{meurtre}) = \textit{crime}$;

(b) synonymes plus spécifiques (sens de $y$ inclut le sens de $x$) – $\textbf{Syn}_{\supset}$
$\textbf{Syn}_{\supset}(\textit{meurtre}) = \textit{assassinat}$;

(c) synonymes à intersection – $\textbf{Syn}_{\cap}$
$\textbf{Syn}_{\cap}(\textit{meurtre}) = \ulcorner \textit{mise à mort} \urcorner$.

FORMALISATION Les coins relevés $\ulcorner \ldots \urcorner$ sont utilisés pour indiquer que le syntagme qu'ils encadrent est une locution (⇒ une unité lexicale) ; dans ce cas-ci, la locution française $\ulcorner$MISE À MORT$\urcorner$.

Aux trois fonctions lexicales de synonymie approximative standard, s'ajoutent les quatre sy-

nonymes approximatifs « de sexe » introduits dans le RL-fr et décrits dans Delaite et Polguère (2013) [1] :

(d) synonymes moins spécifiques vis-à-vis du sexe ($y$ peut être neutre vis-à-vis du sexe, notamment au pluriel) – $\mathbf{Syn}_{\subset}^{\mathbf{sex}}$
$\mathbf{Syn}_{\subset}^{\mathbf{sex}}(avocate) = avocat$;

(e) synonymes plus spécifiques vis-à-vis du sexe – $\mathbf{Syn}_{\supset}^{\mathbf{sex}}$
$\mathbf{Syn}_{\supset}^{\mathbf{sex}}(avocat) = avocate$;

(f) synonymes à intersection donnant le correspondant exact de sexe masculin – **Masc**
**Masc**(*jument*) = *étalon*;

(g) synonymes à intersection donnant le correspondant exact de sexe féminin – **Fem**
**Fem**(*étalon*) = *jument*.

2. $\mathbf{Conv_{ij}}$ : conversif

| **PDD** $x$ | V | N | Adj | Adv |
|---|---|---|---|---|
| **PDD** $y$ | V | N | Adj | Adv |

Lexie fonctionnant comme paraphrase de $x$ moyennant une inversion des actants
$\mathbf{Conv_{21}}$(*inclure*) = *appartenir*, ⌜*faire partie*⌝
$\mathbf{Conv_{3214}}$(*acheter*) = *vendre*

3. **Anti** : antonyme

| **PDD** $x$ | V | N | Adj | Adv |
|---|---|---|---|---|
| **PDD** $y$ | V | N | Adj | Adv |

Lexie ayant un sens opposé à celui de $x$
**Anti**(*respect*) = *irrespect*
**Anti**(*petit*) = *grand*

### 2.2 Relations apparentées à Syn/Anti

4. **Contr** : contrastif

| **PDD** $x$ | V | N | Adj | Adv |
|---|---|---|---|---|
| **PDD** $y$ | V | N | Adj | Adv |

Lexie mise en opposition contrastive avec $x$, sans en être un antonyme
**Contr**(*terre*) = *mer*
**Contr**(*eau*) = *feu*

5. **Epit** : épithète pléonastique

| **PDD** $x$ | V | N | Adj | Adv |
|---|---|---|---|---|
| **PDD** $y$ | Adv | Adj | Adv | Adv |

Modificateur de $x$ sémantiquement redondant, à valeur stylistique
**Epit**(*océan*) = *immense*
**Epit**(*abîme*) = *profond*

6. **Gener** : terme générique

| **PDD** $x$ | V | N | Adj | Adv |
|---|---|---|---|---|
| **PDD** $y$ | — | N | — | — |

Hyperonyme qui fonctionne dans une des deux structures suivantes impliquant $x$ :
– $N_y$ + $Adj_{\text{dérivé de } x \text{ signifiant 'relatif à x'}}$
– $N_x$, $N_{x'}$... *et autres* $N_y$
**Gener**(*gaz*) = *substance* – cf. syntagme *substance gazeuse*
**Gener**(*armoire*) = *meuble* – cf. syntagme *armoire, chaise... et autres meubles*

7. **Figur** : nom métaphorique (figuratif)

| **PDD** $x$ | V | N | Adj | Adv |
|---|---|---|---|---|
| **PDD** $y$ | — | N | — | — |

Le nom $y$ se combine avec $x$ pour former un syntagme qui est une paraphrase de $x$ seul ; c'est-à-dire que $y$ n'ajoute pas de contenu sémantique significatif au sens de $x$
**Figur**(*jalousie*) = *démon* [*de la* ∼]
**Figur**(*honte*) = *rouge* [*de la* ∼].

### 2.3 Dérivations syntaxiques « pures »

8. $\mathbf{S_0}$ : nominalisation

| **PDD** $x$ | V | N | Adj | Adv |
|---|---|---|---|---|
| **PDD** $y$ | N | — | N | N |

Lexie nominale ayant la même valeur sémantique que $x$
$\mathbf{S_0}$(*présenter*) = *présentation*
$\mathbf{S_0}$(*partir*) = *départ*
Dans le cas d'une connexion entre une lexie adjectivale $L_{Adj}$ et le substantif qui signifie 'fait d'être $L_{Adj}$', on utilise la fonction lexicale complexe $\mathbf{S_0Pred}$, où **Pred** est la fonction lexicale standard simple qui dénote la copule (*être*) :
$\mathbf{S_0Pred}$(*obscur*) = *osbscurité*.

9. $\mathbf{V_0}$ : verbalisation

| **PDD** $x$ | V | N | Adj | Adv |
|---|---|---|---|---|
| **PDD** $y$ | — | V | V | V |

Lexie verbale ayant la même valeur sémantique que $x$
$\mathbf{V_0}$(*présentation*) = *présenter*
$\mathbf{V_0}$(*serment*) = *jurer*

10. $\mathbf{A_0}$ : adjectivisation

| **PDD** $x$ | V | N | Adj | Adv |
|---|---|---|---|---|
| **PDD** $y$ | Adj | Adj | — | Adj |

Lexie adjectivale ayant la même valeur sémantique que $x$
$\mathbf{A_0}$(*rotation*) = *giratoire*

1. Il s'agit de fonction lexicales simples proposées, mais dont le statut à l'heure actuel n'est pas « officialisé » hors du projet du RL-fr.

11. **$\mathbf{Adv_0}$** : adverbialisation

| **PDD** $x$ | V | N | Adj | Adv |
|---|---|---|---|---|
| **PDD** $y$ | Adv | Adv | Adv | — |

Lexie adverbiale ayant la même valeur sémantique que $x$
**$\mathbf{Adv_0}$**(*rapide*) = *rapidement*

### 2.4 Dérivés sémantiques nominaux actanciels

12. **$\mathbf{S_i}$** : nom typique de l'actant $i$

| **PDD** $x$ | V | N | Adj | Adv |
|---|---|---|---|---|
| **PDD** $y$ | N | N | N | N |

Lexie nominale servant à dénoter l'actant 1, 2, 3… de $x$
**$\mathbf{S_1}$**(*dire*) = *locuteur*
**$\mathbf{S_2}$**(*dire*) = *paroles*, *propos*…
**$\mathbf{S_3}$**(*dire*) = *destinataire*
**$\mathbf{S_4}$**(*vendre*) = *montant*, *prix*
**$\mathbf{S_i}$** s'emploie fréquemment en conjonction avec l'exposant **usual** pour dénoter le participant actanciel « habituel » ; par exemple :
**$\mathbf{S_1}^{\mathbf{usual}}$**(*angoisse*) = *angoissé*$_{\text{N}}$
**$\mathbf{S_1}^{\mathbf{usual}}$**(*cultiver*) = *cultivateur*

### 2.5 Dérivés sémantiques nominaux circonstanciels

13. **$\mathbf{S_{instr}}$** : nom typique d'instrument

| **PDD** $x$ | V | N | Adj | Adv |
|---|---|---|---|---|
| **PDD** $y$ | N | N | — | — |

Lexie nominale servant à dénoter un circonstant instrumental du fait dénoté par $x$
**$\mathbf{S_{instr}}$**(*gifler*) = *main*
**$\mathbf{S_{instr}}$**(*piqure*) = *seringue*
**$\mathbf{S_{instr}}$**(*crime*) = *arme* [*du* ∼]

14. **$\mathbf{S_{loc}}$** : nom typique de lieu

| **PDD** $x$ | V | N | Adj | Adv |
|---|---|---|---|---|
| **PDD** $y$ | N | N | — | — |

Lexie nominale servant à dénoter un circonstant locatif du fait dénoté par $x$
**$\mathbf{S_{loc}}$**(*bataille*) = ⌜*champ de* ∼⌝
**$\mathbf{S_{loc}}$**(*fumer*) = *fumoir*; *espace fumeur*(*s*)

15. **$\mathbf{S_{med}}$** : nom typique de moyen

| **PDD** $x$ | V | N | Adj | Adv |
|---|---|---|---|---|
| **PDD** $y$ | N | N | — | — |

Lexie nominale servant à dénoter un circonstant de moyen du fait dénoté par $x$
**$\mathbf{S_{med}}$**(*coller*) = *colle*

16. **$\mathbf{S_{res}}$** : nom typique de résultat

| **PDD** $x$ | V | N | Adj | Adv |
|---|---|---|---|---|
| **PDD** $y$ | N | N | — | — |

Lexie nominale servant à dénoter le résultat du fait dénoté par $x$ (= circonstant résultatif)
**$\mathbf{S_{res}}$**(*travailler*) = *fruit* [*de* ART ∼]

### 2.6 Dérivés sémantiques nominaux de type « paramètres »

17. **Sing** : singulatif

| **PDD** $x$ | V | N | Adj | Adv |
|---|---|---|---|---|
| **PDD** $y$ | N | N | — | — |

Lexie nominale servant à dénoter une unité de ce qui est dénoté par $x$
**Sing**(*riz*) = *grain* [*de* ∼]
**Sing**(*pluie*) = *goutte* [*de* ∼]

18. **Mult** : collectif

| **PDD** $x$ | V | N | Adj | Adv |
|---|---|---|---|---|
| **PDD** $y$ | N | N | — | — |

Lexie nominale servant à dénoter un ensemble de ce qui est dénoté par $x$
**Mult**(*chien*) = *meute* [*de* ∼*s*]
**Mult**(*poisson*) = *banc* [*de* ∼*s*]

19. **Cap** : nom de chef

| **PDD** $x$ | V | N | Adj | Adv |
|---|---|---|---|---|
| **PDD** $y$ | N | N | — | — |

Lexie nominale servant à dénoter le responsable/chef de ce qui est dénoté par $x$
**Cap**(*université*) = *président*, **Belg. Québ.** *recteur*
**Cap**(*avion*) = *commandant* (*de bord*)

20. **Equip** : nom d'équipe

| **PDD** $x$ | V | N | Adj | Adv |
|---|---|---|---|---|
| **PDD** $y$ | N | N | — | — |

Lexie nominale servant à dénoter un ensemble de personnes fonctionnant dans le contexte de ce qui est dénoté par $x$
**Equip**(*théâtre*) = *troupe* [*de* ∼]

21. **Germ** : nom de point de « point d'origine »

| **PDD** $x$ | V | N | Adj | Adv |
|---|---|---|---|---|
| **PDD** $y$ | N | N | — | — |

Lexie nominale servant à dénoter le germe, l'origine de ce qui est dénoté par $x$
**Germ**(*colère*) = *ferment*, *levain* [*de la* ∼]

22. **Centr** : nom de centre

| **PDD** $x$ | V | N | Adj | Adv |
|---|---|---|---|---|
| **PDD** $y$ | N | N | — | — |

Lexie nominale servant dénoter le centre ou ce qui est au milieu de ce qui est dénoté par $x$
**Centr**(*problème*) = *cœur*, *noyau* [*du* ∼]
**Centr**(*rue*) = *milieu* [*de la* ∼]

23. **Culm** : nom de point culminant

| **PDD** $x$ | V | N | Adj | Adv |
|---|---|---|---|---|
| **PDD** $y$ | N | N | — | — |

Lexie nominale servant dénoter à la culmination, l'état maximal du fait dénoté par $x$
**Culm**(*colère*) = *paroxysme* [*de la* ∼]
**Culm**(*gloire*) = *sommet* [*de la* ∼]
**Culm**(*joie*) = *comble* [*de la* ∼]

### 2.7 Dérivés sémantiques adjectivaux actanciels

24. **$A_i$** : adjectif dérivé actanciel « pur » pour l'actant $i$ de la lexie $x$, signifiant 'tel qu'il est $i$ de $x$'

| **PDD** $x$ | V | N | Adj | Adv |
|---|---|---|---|---|
| **PDD** $y$ | Adj | Adj | — | — |

Lexie adjectivale servant à dénoter la propriété d'être le participant 1, 2, 3... du fait dénoté par $x$
**$A_1$**(*victoire*) = *victorieux*
**$A_2$**(*victoire*) = $vaincu_{\text{Adj}}$
**$A_3$**(*débat*) = *en* [∼]

25. **$Able_i$** : adjectif dérivé actanciel « de potentiel » pour l'actant $i$ de la lexie $x$, signifiant 'tel qu'il peut être, qu'on peut le rendre, etc. $i$ de $x$' — $y$ est normalement aussi un dérivé morphologique de $x$

| **PDD** $x$ | V | N | Adj | Adv |
|---|---|---|---|---|
| **PDD** $y$ | Adj | Adj | — | — |

Lexie adjectivale servant à dénoter la propriété d'avoir le potentiel d'être le participant 1, 2, 3... du fait dénoté par $x$
**$Able_1$**(*peur*) = $peureux_{\text{Adj}}$
**$Able_1$**(*nuire*) = *nocif*
**$Able_2$**(*lire*) = *lisible*

26. **$Qual_i$** : adjectif dérivé actanciel « de participant virtuel » pour l'actant $i$ de la lexie $x$, signifiant 'tel qu'il est probable qu'il soit, qu'on le rende, etc. $i$ de $x$' — contrairement au cas de **$Able_i$**, $y$ n'est normalement pas un dérivé morphologique de $x$

| **PDD** $x$ | V | N | Adj | Adv |
|---|---|---|---|---|
| **PDD** $y$ | Adj | Adj | — | — |

Lexie adjectivale servant à dénoter la propriété d'avoir une certaine tendance à être le participant 1, 2, 3... du fait dénoté par $x$
**$Qual_1$**(*tromper*) = *malhonnête*
**$Qual_2$**(*tromper*) = *naïf*
Comparer avec les liens **$Able_1$** et **$Able_2$** correspondants :
**$Able_1$**(*tromper*) = *trompeur*
**$Able_2$**(*tromper*) = *trompable*

## 3 Fonctions lexicales syntagmatiques

### 3.1 Collocatifs de type modificateurs

27. **Magn** : intensificateur

| **PDD** $x$ | V | N | Adj | Adv |
|---|---|---|---|---|
| **PDD** $y$ | Adv | Adj | Adv | Adv |

Modificateur adjectival ou adverbial de $x$ qui exprime un sens du type 'très', 'intense', 'beaucoup', etc.
**Magn**(*amour*) = *fou*
**Magn**(*peur*) = *bleue*
**Magn**(*boire*) = ⌜*comme un trou*⌝
**Magn**(*gagner*) = ⌜*haut la main*⌝
**Magn**($mort_{\text{Adj}}$) = *raide*
**Magn**(*enceinte*) = ⌜*jusqu'aux yeux*⌝
Noter la possibilité de *valeurs fusionnées* : elles expriment à la fois la base et le collocatif ; une valeur fusionnée est donc une paraphrase de la collocation au complet. La présence d'une telle valeur est indiquée par le symbole « // ». Par exemple :
**Magn**(*défaite*) =

> *grande*, *grosse*, *lourde*, *grave*, *sévère*, *sérieuse* < *écrasante*, *terrible* < *complète*, *totale* //*débâcle* //*déroute* //**fam** *raclée*[2]

28. **Ver** : confirmateur

| **PDD** $x$ | V | N | Adj | Adv |
|---|---|---|---|---|
| **PDD** $y$ | Adv | Adj | Adv | Adv |

Modificateur adjectival ou adverbial de $x$ qui exprime un sens du type 'tel qu'il faut', 'tel que ça doit être', etc.
**Ver**(*conseil*) = *précieux*, *utile*
**Ver**(*succès*) = *mérité*
**Ver**(*regretter*) = *sincèrement*

29. **Bon** : laudatif

| **PDD** $x$ | V | N | Adj | Adv |
|---|---|---|---|---|
| **PDD** $y$ | Adv | Adj | Adv | Adv |

Modificateur adjectival ou adverbial de $x$ qui exprime un sens du type 'bien', 'bon', etc.
**Bon**(*conseil*) = *bon*
**Bon**(*choix*) = *heureux*
**Bon**(*temps*) = *magnifique*

2. L'indication **fam** est une marque d'usage spécifiant que la lexie en question relève du registre familier.

### 3.2 FL complexes

Les FL simples peuvent se combiner pour former des *FL complexes*. Notamment, la FL simple paradigmatique **Anti** se combine avec les FL simples syntagmatiques **Magn**, **Ver** et **Bon** pour former les FL complexes :

- **AntiMagn**
- **AntiVer**
- **AntiBon**

Par exemple :

**AntiMagn**(*salaire*) = ⌜*de misère*⌝
**AntiVer**(*accusation*) = *fausse*
**AntiBon**(*temps*) = ⌜*de chien*⌝

### 3.3 Collocatifs verbaux

#### 3.3.1 Verbes supports

Collocatifs verbaux d'une base nominale qui fonctionne comme sujet ou complément de ces verbes. Ceux-ci sont sémantiquement vides ou possèdent un sens déjà contenu dans celui de la base nominale. Ils servent en fait à « verbaliser » un prédicat nominal.

Il existe trois types de verbes supports, distingués par la position syntaxique qu'occupe la base relativement au collocatif.

30. **Oper$_i$** : verbe support prenant $x$ comme premier complément et l'expression de l'actant $i$ de $x$ comme sujet grammatical

| **PDD** $x$ | V | N | Adj | Adv |
|---|---|---|---|---|
| **PDD** $y$ | — | V | — | — |

Il s'agit du type le plus fréquent de verbe support.

**Oper$_1$**(*suprématie*) = *détenir* [ART ∼]
**Oper$_1$**(*remarque*) = *faire* [ART ∼]
**Oper$_1$**(*méfait*) = *perpétrer* [ART ∼]
**Oper$_2$**(*danger*) = *courir* [ART ∼]
**Oper$_2$**(*applaudissements*) = *recueillir* [ART ∼]

On observe les faits suivants :

- « **Oper$_1$**($x$) + $x$ » est une paraphrase de « **V$_0$**($x$) »

  (1) a. *Luce a poussé un cri.*
  b. *Luce a crié.*

- **Oper$_2$** entretient la même relation de paraphrase (conversive) avec **Oper$_1$** que l'on trouve entre la voix passive et la voix active.

  (2) a. *Luce fait*$_{[= \textbf{Oper}_1]}$ *un compliment à Félix.*
  ≡
  *Luce complimente*$_{[\text{voix active}]}$ *Félix.*
  b. *Félix reçoit*$_{[= \textbf{Oper}_2]}$ *un compliment de Luce.*
  ≡
  *Félix est complimenté*$_{[\text{voix passive}]}$ *par Luce.*

31. **Func$_i$** : verbe support prenant $x$ comme sujet grammatical et l'expression de l'actant $i$ de $x$ comme premier complément ; **Func$_0$** (c'est-à-dire, $i = 0$) signifie que le verbe support ne prend pas de complément exprimant un actant de $x$

| **PDD** $x$ | V | N | Adj | Adv |
|---|---|---|---|---|
| **PDD** $y$ | — | V | — | — |

Il s'agit d'un verbe support conversif de **Oper$_i$**

**Func$_0$**(*pluie*) = *tomber*
**Func$_0$**(*tonnerre*) = *gronder*
**Func$_1$**(*responsabilité*) = *incomber* [*à* N$_X$]
**Func$_2$**(*danger*) = *menacer* [N$_Y$]
**Func$_2$**(*interdiction*) = *frapper* [N$_Y$]

32. **Labor$_{ij}$** : verbe support prenant $x$ comme second complément, l'expression de l'actant $i$ de $x$ comme sujet grammatical et l'expression de l'actant $j$ de $x$ comme premier complément

| **PDD** $x$ | V | N | Adj | Adv |
|---|---|---|---|---|
| **PDD** $y$ | — | V | — | — |

Verbe support qui a pour effet de donner une position communicativement périphérique à l'expression de $x$ dans la phrase, au profit de l'expression des actants $i$ et $j$
**Labor$_{12}$**(*estime*) = *avoir*, *tenir* [N$_Y$ *en* ∼]
**Labor$_{12}$**(*silence*) = *passer* [N$_Y$ *sous* ∼]

On peut clairement voir les différentes structures impliquées par les trois fonctions lexicales de verbes supports en contrastant les applications suivantes :

- **Oper$_1$**(*analyse*) = *faire* [ART ∼]
- **Func$_1$**(*analyse*) = *provenir* [*de* N$_X$]
- **Func$_2$**(*analyse*) = *concerner* [N$_Y$], *traiter* [*de* N$_Y$], *porter* [*sur* N$_Y$]
- **Labor$_{12}$**(*analyse*) = *soumettre* [N$_Y$ *à* ART ∼]

#### 3.3.2 Verbes de réalisation

Il s'agit de trois types de collocatifs verbaux qui sont les pendants des trois types de verbes supports vus précédemment. Contrairement aux verbes supports, ils apportent dans la collocation un contenu sémantique : celui de 'réaliser', 'utiliser', 'accomplir', etc., au sens le plus large.

33. **Real$_i$** : verbe de réalisation prenant $x$ comme premier complément et l'expression de l'actant $i$ de $x$ comme sujet grammatical → « verbe de réalisation » contrôlant la même configuration syntaxique que les verbes supports du type **Oper$_i$**

| **PDD** $x$ | V | N | Adj | Adv |
|---|---|---|---|---|
| **PDD** $y$ | — | V | — | — |

Signifie ‘utiliser’, ‘réaliser’, etc.

**Real$_1$**(*peine*) = *infliger* [ART ∼]
**Real$_2$**(*peine*) = *purger* [ART ∼]
**Real$_1$**(*avion*) = *piloter* [ART ∼]

34. **Fact$_i$** : verbe de réalisation prenant $x$ comme sujet grammatical et l'expression de l'actant $i$ de $x$ comme premier complément → « verbe de réalisation » contrôlant la même configuration syntaxique que les verbes supports de type **Func$_i$** ; **Fact$_0$** (c'est-à-dire, $i = 0$) signifie que le collocatif verbal ne prend pas de complément exprimant un actant de $x$

| **PDD** $x$ | V | N | Adj | Adv |
|---|---|---|---|---|
| **PDD** $y$ | — | V | — | — |

Verbe de réalisation conversif de **Real$_i$** ; signifie ‘fonctionner’, ‘faire ce qui est attendu’, ‘avoir un effet’, etc.

**Fact$_0$**(*rêve*) = *se réaliser*
**Fact$_1$**(*effroi*) = *paraylser* [N$_X$]
**Fact$_2$**(*insolence*) = *choquer*, *offenser*, *offusquer* [N$_Y$]

35. **Labreal$_{ij}$** : verbe de réalisation prenant $x$ comme second complément, l'expression de l'actant $i$ de $x$ comme sujet grammatical et l'expression de l'actant $j$ de $x$ comme premier complément → « verbe de réalisation » contrôlant la même configuration syntaxique que les verbes supports de type **Labor$_{ij}$**

| **PDD** $x$ | V | N | Adj | Adv |
|---|---|---|---|---|
| **PDD** $y$ | — | V | — | — |

Verbe de réalisation qui a pour effet de donner une position communicativement périphérique à l'expression de $x$ dans la phrase, au profit de l'expression des actants $i$ et $j$

**Labreal$_{12}$**(*piège*) = *prendre* [N$_Y$ *dans* ART ∼]
**Labreal$_{12}$**(*imagination*) = *voir* [N$_Y$ *en* ∼]

## Références

Delaite, C. et Polguère, A. (2013). Sex-Based Nominal Pairs in the French Lexical Network: It's Not What You Think. Dans *Proceedings of the 6th International Conference on Meaning-Text Theory (MTT'13)*, pages 29–40, Prague, Tchèque, République.

Mel'čuk, I., Clas, A. et Polguère, A. (1995). *Introduction à la lexicologie explicative et combinatoire*. Duculot, Paris/Louvain-la-Neuve.

# Bibliographie


Agnar Aamodt et Enric Plaza. Case-based reasoning ; foundational issues, methodological variations, and system approaches. *AI COMMUNICATIONS*, 7(1) :39–59, 1994. pages 2

István Albert et Réka Albert. Conserved network motifs allow protein–protein interaction prediction, 2004. pages 135

Sue B. T. Atkins. Bilingual dictionaries : Past, present and future. Dans Martin Gellerstam, Jerker Järborg, Sven-Göran Malmgren, Kerstin Norén, Lena Rogström et Catalina Röjder Papmehl, dir., *Proceedings of the 7th EURALEX International Congress*, pages 515–546, Göteborg, Sweden, aug 1996. Novum Grafiska AB. pages 1, 39

Marco Baroni, Silvia Bernardini, Adriano Ferraresi et Eros Zanchetta. The wacky wide web : a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3) :209–226, 2009. pages 7, 70

Lucie Barque. *Description et formalisation de la polysémie régulière du français*. Thèse de doctorat, Université Paris Diderot - Paris 7, 2008. pages 57, 126, 136, 189, 193

Lucie Barque et François-Régis Chaumartin. Regular Polysemy in WordNet. *JLCL - Journal for Language Technology and Computational Linguistics*, 24(2) :5–18, 2009. pages 57, 136

Ahmed Belderrar. *Extraction des sous-graphes : identification des microarchitectures dans les logiciels évolutifs orientés objets*. masters, École Polytechnique de Montréal, août 2011. pages 137

Bélaa Bollobás et Oliver Riordan. The diameter of a scale-free random graph. *Combinatorica*, 24(1) :5–34, janvier 2004. pages 83

Vincent Bontems. L'analogie dans l'épistémologie historique de ferdinand gonseth : Les concepts post-phénoménologiques de schéma, horizon de réalité et référentiel. *Bulletin d'Analyse Phénoménologique*, 3(3), novembre 2007. pages 16

Stephen P. Borgatti, Martin G. Everett et Paul R. Shirey. LS sets, lambda sets and other cohesive subsets. *Social Networks*, 12(4) :337–357, décembre 1990. pages 134

Armelle Boussidan, Sylvain Lupone et Sabine Ploux. La malbouffe : un cas de néologie et de glissement sémantique fulgurants. *Du thème au terme, émergence et*

*lexicalisation des connaissances", Toulouse, France. 8 ème conférence internationale Terminologie et Intelligence Artificielle*, 2009. pages 194

Ulrik Brandes, Markus Eiglsperger, Ivan Herman, Michael Himsolt et M.Scott Marshall. Graphml progress report structural layer proposal. Dans Petra Mutzel, Michael Jünger et Sebastian Leipert, dir., *Graph Drawing*, volume 2265 de *Lecture Notes in Computer Science*, pages 501–512. Springer Berlin Heidelberg, 2002. pages 137

Coen Bron et Joep Kerbosch. Algorithm 457 : Finding all cliques of an undirected graph. *Commun. ACM*, 16(9) :575–577, septembre 1973. pages 134

Vincent Claveau et Marie-Claude L'Homme. Apprentissage par analogie pour la structuration de terminologie - utilisation comparée de ressources endogènes et exogènes. Dans *Terminologie et Intelligence Artificielle, TIA'05*, Rouen, France, April 2005. pages 26, 28

Luigi P. Cordella, Pasquale Foggia, Carlo Sansone et Mario Vento. An improved algorithm for matching large graphs. Dans *3rd IAPR-TC15 Workshop on Graph-based Representations in Pattern Recognition, Cuen*, page 149–159, 2001. pages 141

Georgette Dal. Analogie et lexique construit : quelles preuves ? Dans *Cahiers de grammaire*, pages 9–30. Université de Toulouse-le-Mirail, 2003. pages 23, 24

Jean-Paul Delahaye. Que le monde est petit ! *Pour la science*, 308 :98–103, 2003. pages 79, 81

Candice Delaite et Alain Polguère. Sex-Based Nominal Pairs in the French Lexical Network : It's Not What You Think. Dans *Proceedings of the 6th International Conference on Meaning-Text Theory (MTT'13)*, pages 29–40, Prague, Tchèque, République, 2013. pages 47, 113

Gilles Deleuze. Anti-œdipe et mille plateaux. Transcription du cours donné à Vincennes le 14 janvier 1974, 1974. pages 7

Yann Desalle. *Réseaux lexicaux, métaphore, acquisition : une approche interdisciplinaire et inter-linguistique du lexique verbal*. Thèse de doctorat, Université Toulouse le Mirail - Toulouse II, mai 2012. pages 36, 42

Yann Desalle, Bruno Gaume, Karine Duvignau, Hintat Cheung, Shu-Kai Hsieh, Pierre Magistry et Jean-Luc Nespoulous. Skillex, an action labelling efficiency score : the case for french and mandarin. *Proceedings of the 36th Annual Conference of the Cognitive Science Society*, pages 409–414, 2014a. pages 42

Yann Desalle, Emmanuel Navarro, Yannick Chudy, Pierre Magistry et Bruno Gaume. Bacanal : Balades aléatoires courtes pour analyses lexicales, application à la substitution lexicale. Dans *Actes de TALN 2014*, pages 206–217, Marseille, France, juillet 2014b. pages 42

Patrick Drouin, Annie Paquin et Nathan Ménard. Extraction semi-automatique des néologismes dans la terminologie du terrorisme. *Actes des 8e Journées internationales d'Analyse statistique des Données Textuelles (JADT 2006)*, pages 389–400, 2006. pages 194

Karine Duvignau et Bruno Gaume. Linguistic, psycholinguistic, and computational approaches to the lexicon : for early verb-learning based on analogy. *Cognitive Systems*, 6(2/3) :255–269, 2004. pages 36, 42

Brian Falkenhainer, Kenneth D. Forbus et Dedre Gentner. The structure-mapping engine : Algorithm and examples. *Artificial Intelligence*, 41(1) :1–63, novembre 1989. pages 15, 17, 18, 20

Christiane Fellbaum. *WordNet : an electronic lexical database*. Language, Speech and Communication. MIT Press, 1998. pages 33, 39, 192

Nabil Gader, Aurore Koehl et Alain Polguère. A Lexical Network with a Morphological Model in It. Dans *Proceedings of the 4th Workshop on Cognitive Aspects of the Lexicon (CogALex)*, pages 154–165, Dublin, Irlande, août 2014a. Association for Computational Linguistics and Dublin City University. pages 62

Nabil Gader, Veronika Lux-Pogodalla et Alain Polguère. Hand-Crafting a Lexical Network With a Knowledge-Based Graph Editor. Dans *Proceedings of the Third Workshop on Cognitive Aspects of the Lexicon (CogALex III)*, pages 109–125, Mumbai, 15 décembre 2012. The COLING 2012 Organizing Committee. pages 39, 58

Nabil Gader, Sandrine Ollinger et Alain Polguère. One Lexicon, Two Structures : So What Gives ? Dans Piek Vossen Heili Orav, Christiane Fellbaum, dir., *Proceedings of the Seventh Global Wordnet Conference (GWC2014)*, pages 163–171, Tartu, Estonie, 2014b. Global WordNet Association. pages 40

Benoit Gaillard, Bruno Gaume et Emmanuel Navarro. Invariants and variability of synonymy networks : Self mediated agreement by confluence. Dans *Proc. of TextGraphs-6 : Graph-based Methods for NLP*, pages 15–23, Portland, june 2011. ACL. pages 41, 73

Bruno Gaume. Analogie et proxémie dans les réseaux petits mondes. *Revue d'Intelligence Artificielle*, 17(5-6) :935–951, 2003. pages 36

Bruno Gaume. Balades aléatoires dans les petits mondes lexicaux. *Information interaction intelligence*, 4(2) :39–96, 2004. pages 25, 41, 42, 79, 80, 195

Dedre Gentner. Structure-mapping : A theoretical framework for analogy. *Cognitive Science*, 7(2) :155–170, 1983. pages 2, 15, 16, 17, 30, 57, 94, 141

Dedre Gentner et Keith J. Holyoak. Reasoning and learning by analogy : Introduction. *American Psychologist*, 52 :32–34, 1997. pages 15, 16

Dedre Gentner et Arthur B. Markman. Structure mapping in analogy and similarity. *American Psychologist*, 52 :45–56, 1997. pages 15, 16, 18, 19, 20

Karen González Orellana. Creación de una Red Léxica del Español para el análisis comparativo de los grafos léxicos del español y del francés. V Congreso Internacional de Lexicografía Hispánica, 2012. pages 40

Vannina Goossens. *Propositions pour le traitement de la polysémie régulière des noms d'affect*. Thèse de doctorat, Université de Grenoble, 2011. pages 57, 136

Joseph E. Grimes. Inverse lexical functions. Dans *Meaning-text theory : Linguistics, lexicography, and implications*, pages 350–364. University of Ottawa Press, James Steele édition, 1990. pages 77

Benoît Habert. *Instruments Et Ressources Électroniques Pour Le Français*. Ophrys, 2005. pages 2

Nabil Hathout. Acquisition morphologique à partir d'un dictionnaire informatisé. Dans *Actes de TALN 2009*, Senlis, 2009. ATALA. pages 24, 25, 26, 153, 155

Nabil Hathout. Morphonette : a paradigm-based morphological network. *Lingue e linguaggio*, 2011(2) :243–262, 2011. pages 24, 25, 26, 153, 155

Nabil Hathout et Fiammetta Namer. La base lexicale démonette : entre sémantique constructionnelle et morphologie dérivationnelle. Dans *Actes de TALN 2014*, pages 208–219, Marseille, 2014. pages 26

Serge Heiden, Jean-Philippe Magué et Bénédicte Pincemin. TXM : Une plateforme logicielle open-source pour la textométrie - conception et développement. Dans Luca Giuliano Sergio Bolasco, Isabella Chiari, dir., *Statistical Analysis of Textual Data - Proceedings of 10th International Conference Journées d'Analyse statistique des Données Textuelles*, volume 2, pages 1021–1032, Rome, Italie, juin 2010. Edizioni Universitarie di Lettere Economia Diritto. pages 71, 190

Didier Henrion. *Les quinze livres des éléments géométriques d'Euclide : plus le livre des donnez du mesme Euclide aussi traduict en françois par ledit Henrion, et imprimé de son vivant / Traduicts en françois par D. Henrion...* Veuve Henrion (Paris), 1632. pages 9, 10

Paul Hermann. *Prinzipien der Sprachgeschichte*. Max Niemeyer, 1960. pages 12

Élodie Jactel. *Le RLF pour tous. Paraphrasage des liens de fonctions lexicales dans le Réseau Lexical du Français*. Mémoire de Master Européen de Lexicographie (European Master in Lexicography, EMLex), Université de Lorraine, june 2013. pages 46

Roman Jakobson. *Essais de linguistique générale*. Minuit, Paris, 1963. pages 87, 126, 128

Anne-Laure Jousse. *Modèle de structuration des relations lexicales fondé sur le formalisme des fonctions lexicales*. Thèse ou mémoire numérique / electronic thesis or dissertation, Université de Montréal et Paris Diderot (Paris 7), avril 2010. pages 46, 48

Mi Hyun Kim. Definition of body element nouns in a Korean Lexical Network. Dans Deny Kwary, dir., *Proceedings of AsiaLex 2013*. Airlangga University Press, 2013. pages 40

Philipp Koehn et Christof Monz. Manual and automatic evaluation of machine translation between european languages. Dans *Proceedings on the Workshop on Statistical Machine Translation*, pages 102–121, New York City, June 2006. Association for Computational Linguistics. pages 29

JanetL. Kolodner. An introduction to case-based reasoning. *Artificial Intelligence Review*, 6(1) :3–34, 1992. pages 27

Mathieu Lafourcade et Alain Joubert. Détermination des sens d'usage dans un réseau lexical construit grâce à un jeu en ligne. Dans *Actes de TALN 2008*, pages 189–199, Avignon, France, juin 2008. pages 40, 42, 191

Mathieu Lafourcade et Alain Joubert. Computing trees of named word usages from a crowdsourced lexical network. *Investigationes Linguisticae*, 21 :39–56, 2010. pages 40, 191

George Lakoff. *Women, Fire and Dangerous Things : What Categories Reveal About the Mind*. University of Chicago Press, Chicago, 1987. pages 34

Philippe Langlais et Alexandre Patry. Enrichissement d'un lexique bilingue par apprentissage analogique. *Traitement Automatique des Langues (TAL)*, 49 (varia) : 13–40, 2008. pages 28, 29

Philippe Langlais et François Yvon. Issues in analogical inference over sequences of symbols : A case study on proper name transliteration. Dans Henri Prades et Gilles Richard, dir., *Computational Approaches to Analogical Reasoning : Current Trends*, pages 59–82. Springer-Verlag Berlin Heidelberg, 2014. pages 2

Yves Lepage. Défense et illustration de l'analogie. Dans *Actes de TALN 2001*, Tours, juillet 2001. pages 23, 24, 87

Yves Lepage. *De l'analogie rendant compte de la commutation en linguistique*. Thèse de doctorat, Université Joseph-Fourier - Grenoble I, mai 2003. pages 2, 9, 11, 12, 13, 14, 15, 19, 20, 21, 22, 23, 28, 32, 35, 54, 55, 87, 88, 89, 90, 91, 106, 141, 149

Yves Lepage. Analogie en traitement automatique des langues. application à la traduction automatique. Dans *Actes de TALN 2006*, pages 781–791, Leuven, Belgium, Avril 2006. pages 11, 23

Yves Lepage. Analogies between binary images : Application to chinese characters. Dans Henri Prade et Gilles Richard, dir., *Computational Approaches to Analogical Reasoning : Current Trends*, numéro 548 dans Studies in Computational Intelligence, pages 25–57. Springer Berlin Heidelberg, janvier 2014. pages 11, 23

Vladimir Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8) :707–710, 1966. pages 88

Geoffrey Ernest Richard Lloyd. *Polarity and Analogy : Two Types of Argumentation in Early Greek Thought*. Bristol Classical Press, 1966. pages 16

Sylvain Loiseau, Philippe Gréa et Jean-Philippe Magué. Dictionnaires, théorie des graphes et structures lexicales. *Revue de Sémantique et de Pragmatique*, 27 :51–78, 2011. pages 41

Veronika Lux-Pogodalla. Intégration relationnelle des exemples lexicographique dans un réseau lexical. Dans *Actes de TALN 2014*, pages 586–591, Marseille, France, juillet 2014. pages 69

Douglas L. Medin, Robert L. Goldstone et Dedre Gentner. Similarity involving attributes and relations : Judgments of similarity and difference are not inverses. *Psychological Science*, 1(1) :64–69, janvier 1990. pages 30, 57, 90, 94, 141

Igor Mel'čuk. Lexical functions : A tool for the description of lexical relations in the lexicon. Dans Leo Wanner, dir., *Lexical Functions in Lexicography and Natural Language Processing*, volume 31 de *Language Companion Series*, pages 37 – 102. John Benjamins, Amsterdam/Philadelphia, studies in language édition, 1996. pages 41

Igor Mel'čuk. Parties du discours et locutions. *Bulletin de la Société de linguistique de Paris*, 101(1) :29–65, 2006. pages 96, 146

Igor Mel'čuk, André Clas et Alain Polguère. *Introduction à la lexicologie explicative et combinatoire*. 256 pages. Duculot, Paris/Louvain-la-Neuve, 1995. pages 27, 36, 41, 43, 45, 46, 54, 55, 94, 97

Jasmina Milicevic et Alain Polguère. Ambivalence sémantique des noms de communication langagière du français. Dans Institut de Linguistique Française (ILF), dir., *Actes du 2e Congrès Mondial de Linguistique Française (CMLF'10)*, pages 1029–1050, La Nouvelle-Orléans, États-Unis, 2010. pages 66

R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii et U. Alon. Network motifs : simple building blocks of complex networks. *Science*, 298(5594) :824–827, October 2002. pages 135

Ron Milo, Shalev Itzkovitz, Nadav Kashtan, Reuven Levitt, Shai Shen-Orr, Inbal Ayzenshtat, Michal Sheffer et Uri Alon. Superfamilies of evolved and designed networks. *Science*, 303(5663) :1538–1542, 2004. pages 135

Makoto Nagao. A framework of a mechanical translation between japanese and english by analogy principle. Dans *Proc. Of the International NATO Symposium on Artificial and Human Intelligence*, pages 173–180, New York, NY, USA, 1984. Elsevier North-Holland, Inc. pages 19, 20

Emmanuel Navarro. *Métrologie des graphes de terrain, application à la construction de ressources lexicales et à la recherche d'information*. Thèse de doctorat, Institut National Polytechnique de Toulouse - INPT, novembre 2013. pages 79, 81, 134

Emmanuel Navarro, Yannick Chudy et Bruno Gaume. Détection de communautés sur un graphe biparti et application à la classification automatique des résultats d'une recherche web (système Kodex). Journée sur les Graphes pour la Fouille dans le Web, février 2010. pages 134

Roberto Navigli et Simone Paolo Ponzetto. Babelnet : Building a very large multilingual semantic network. Dans *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 216–225, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics. pages 39

M. E. J. Newman. The structure of scientific collaboration networks. *Proceedings of the National Academy of Sciences of the United States of America*, 98(2) :404–409, January 2001. pages 81, 82

M. E. J. Newman. The structure and function of complex networks. *SIAM REVIEW*, 45 :167–256, 2003. pages 79, 81, 82

Sandrine Ollinger. Regroupement de structures de dérivations lexicales par raisonnement analogique. Dans Brigitte Bigi, dir., *Actes de Récital 2014*, pages 92–103, Marseille, France, 2014. pages 131

Sandrine Ollinger et Mathieu Valette. la créativité lexicale : des pratiques sociales aux textes. Dans *CINEO'08*, pages 25–40, Barcelone, Spain, mai 2008. pages 194

Marie-Sophie Pausé. *Modélisation de la structure lexico-syntaxique des locutions : approche lexicographique*. Mémoire de Master en Sciences du Langage et Didactique des Langues spécialité Lexique, Textes, Discours, Université de Lorraine, june 2014. pages 53

Tiago P. Peixoto. The graph-tool python library. *figshare*, 2014. pages 137, 164

Vito Pirrelli et Francois Yvon. The hidden dimension : a paradigmatic view of data-driven nlp. *Journal of Experimental & Theoretical Artificial Intelligence*, 11(3) : 391–408, 1999. pages 23, 28

Sabine Ploux et Bernard Victorri. Construction d'espaces sémantiques à l'aide de dictionnaires de synonymes. *Traitement Automatique des Langues*, 39 :161–182, 1998. pages 42, 134

Alain Polguère. Lexical function standardness. Dans Léo Wanner, dir., *Selected Lexical and Grammatical Issues in the Meaning-text Theory : In Honour of Igor Mel čuk*, pages 43–95. John Benjamins Publishing, Amsterdam/Philadelphia, 2007. pages 47

Alain Polguère. *Lexicologie et sémantique lexicale : Notions fondamentales*. Paramètres. PUM, Montréal, $2^{ieme}$ édition, 2008. Édition revue et augmentée. pages 44, 55, 93

Alain Polguère. Lexical systems : graph models of natural language lexicons. *Language Resources and Evaluation*, 43(1) :41–55, 2009. pages 1, 39

Alain Polguère. Classification sémantique des lexies fondée sur le paraphrasage. *Cahiers de lexicologie*, 98 :197–211, août 2011a. pages 64, 65

Alain Polguère. Mémo RLF. Étapes dans la rédaction d'une entrée de vocable, 2011b. pages 43

Alain Polguère. From Writing Dictionaries to Weaving Lexical Networks. *International Journal of Lexicography*, 2014a. pages 1, 39

Alain Polguère. Principes de modélisation systémique des réseaux lexicaux. Dans *Actes de TALN 2014*, pages 79–90, Marseille, France, juillet 2014b. pages 1, 39, 41, 54

Alain Polguère et Dorota Sikora. Modèle lexicographique de croissance du vocabulaire fondé sur un processus aléatoire, mais systématique. *Enseigner le lexique*, pages 35–63, 2013. pages 1, 41

Stéfan Popovic. Métalangage de vulgarisation des liens de fonctions lexicales. Dans *Proceedings of The First International Conference on Meaning-Text Theory*, Paris, France, 2013. pages 46

Erzsébet Ravasz et Albert-László Barabási. Hierarchical organization in complex networks. *Phys. Rev. E*, 67 :026112, Feb 2003. pages 83

Coralie Reutenauer, Evelyne Jacquey et Sandrine Ollinger. Neologismes de sens : contribution à leur caracterisation dans un corpus autour du thème de la crise financière. Dans *II Congrès International de Néologie des Langues Romanes (Cineo2011)*, São Paulo, Brazil, décembre 2011. pages 194

Pedro Ribeiro. *Efficient and Scalable Algorithms for Network Motifs Discovery*. Thèse de doctorat, Faculty of Science of the University of Porto, June 2011. pages 137

Ferdinand de Saussure. *Cours de Linguistique Générale*. Payot, 1916. pages 21, 22, 194

Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. Dans *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK, 1994. pages 72

Helmut Schmid. Improvements in part-of-speech tagging with an application to german. Dans *Proceedings of the ACL SIGDAT-Workshop*, pages 47–50, 1995. pages 72

Thierry Selva, Serge Verlinde et Jean Binon. Vers une deuxième génération de dictionnaires électroniques. *TAL. Traitement automatique des langues*, 44(2) : 177–197, 2003. pages 1, 39

Shai S. Shen-Orr, Ron Milo, Shmoolik Mangan et Uri Alon. Network motifs in the transcriptional regulation network of escherichia coli. *Nature Genetics*, 31 : 1061–4036, 2002. pages 138

Mariano Sigman et Guillermo A. Cecchi. Global organization of the wordnet lexicon. *In Proceedings of the National Academy of Sciences*, 99(3) :1742–1747, 2002. pages 79

Dorota Sikora. D'achèvement à état : une affaire de polysémie. 11th International Conference on Actionality, Tense, Aspect, Modality/Evidentiality, 2014. pages 57, 136, 190, 193

Dennis Spohr. *Towards a Multifunctional Lexical Resource*, volume 141 de *Lexicographica. Series Maior*. De Gruyter, 2012. pages 1, 39

Dennis Spohr et Ulrich Heid. Modeling monolingual and bilingual collocation dictionaries in description logics. Dans *Proceedings of the EACL Workshop on Multiwords and Multilinguality*, pages pp. 65 – 72, Trento, Italia, 2006. in connection with EACL-2006. pages 40

Nicolas Stroppa. *Définitions et caractérisations de modèles à base d'analogies pour l'apprentissage automatique des langues naturelles*. Thèse doctorat, Télécom Paris Tech, 2005. pages 2, 21, 22, 23, 28, 91

Nicolas Stroppa et François Yvon. An analogical learner for morphological analysis. Dans *Proceedings of the Ninth Conference on Computational Natural Language Learning (CoNLL'2005)*, pages 120–127, Ann Arbor, MI, 2005. pages 21, 22, 26

Nicolas Stroppa et François Yvon. Du quatrième de proportion comme principe inductif : une proposition et son application à l'apprentissage de la morphologie. *TAL. Traitement automatique des langues*, 47(1), 2006. pages 23

Lionel Tabourier. *Méthode de comparaison des topologies de graphes complexes : applications aux réseaux sociaux*. Thèse de doctorat, Paris 6, 2010. pages 74, 79, 135

Kota Takeya, Jing Sun et Yves Lepage. The number of proportional analogies between marker-based chunks in 11 european languages. Dans *Proceedings of the 17th Annual Meeting of the Association for Natural Language Processing*, pages 677–680, 2011. pages 23

Peter D. Turney. Similarity of semantic relations. *Comput. Linguist.*, 32(3) :379–416, septembre 2006. pages 2, 30, 33, 34, 57, 92, 94, 95, 123, 124, 128, 141

Peter D. Turney. The latent relation mapping engine : Algorithm and experiments, décembre 2008a. pages 19, 33

Peter D. Turney. A uniform approach to analogies, synonyms, antonyms, and associations. Dans *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1*, COLING '08, page 905–912, Stroudsburg, PA, USA, 2008b. Association for Computational Linguistics. pages 20, 32

Peter D. Turney. Domain and function : A dual-space model of semantic relations and compositions. *J. Artif. Int. Res.*, 44(1) :533–585, mai 2012. pages 32

Peter D. Turney. Distributional semantics beyond words : Supervised learning of analogy and paraphrase. *TACL*, 1 :353–366, 2013. pages 32

Peter D. Turney et Michael L. Littman. Corpus-based learning of analogies and semantic relations. *Machine Learning*, 60(1-3) :251–278, août 2005. pages 30

T. Veale. A typology of lexical analogy in WordNet. Dans *Proceedings of the 3rd Global WordNet Conference*, 2006. pages 34

Tony Veale. The analogical thesaurus. Dans John Riedl et Randall W. Hill Jr., dir., *IAAI*, pages 137–142. AAAI, 2003. pages 35

Tony Veale. Wordnet sits the s.a.t. : A knowledge-based approach to lexical analogy. Dans *Proceedings of ECAI'2004, the 16th European Conf. on Artificial Intelligence*. John Wiley : London, 2004. pages 2, 30, 33, 34

Tony Veale et Guofu Li. Analogy as an organizational principle in the construction of large knowledge-bases. Dans Henri Prade et Gilles Richard, dir., *Computational Approaches to Analogical Reasoning : Current Trends*, volume 548 de *Studies in Computational Intelligence*, pages 83–101. Springer Berlin Heidelberg, 2014. pages 20, 35

Robert A. Wagner et Michael J. Fischer. The string-to-string correction problem. *J. ACM*, 21(1) :168–173, janvier 1974. pages 88, 106

Duncan J. Watts et Steven H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684) :440–442, juin 1998. pages 79, 81

Sebastian Wernicke. Efficient detection of network motifs. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 3(4) :347–359, 2006. pages 137

Sebastian Wernicke et Florian Rasche. Fanmod : A tool for fast network motif detection. *Bioinformatics*, 22(9) :1152–1153, mai 2006. pages 137

E. David Wong. Self-generated analogies as a tool for constructing and evaluating explanations of scientific phenomena. *Journal of Research in Science Teaching*, 30 (4) :367–380, avril 1993. pages 16, 19, 188

# Index

## Résumé

La lexicographie contemporaine, en mettant à profit les avancées théoriques et pratiques de l'informatique et de la linguistique, s'est affranchie de l'organisation linéaire imposée par les ouvrages papier. Elle s'est attachée à définir de nouveaux modèles de description et met aujourd'hui à disposition de la communauté des ressources formelles et cohérentes offrant de multiples possibilités d'exploitations automatiques. Cette thèse concentre son attention sur le modèle des systèmes lexicaux proposé par la Lexicographie Explicative et Combinatoire. Plus précisément, elle s'intéresse au Réseau Lexical du Français, en cours de développement. En tant que système lexical, cette ressource est un graphe lexical monolingue. Elle est constituée d'un ensemble de sommets, les unités lexicales du français, entre lesquels sont encodées de nombreuses relations, en grande majorité syntaxico-sémantiques. La présente thèse pose les bases d'une exploration de cette ressource lexicographique par raisonnement analogique. Elle débute par une revue sélective de la formalisation et de l'informatisation de l'analogie en traitement automatique des langues, dans le cas précis de l'étude du lexique. Elle définit ainsi le principe de l'exploration réalisée comme un regroupement de structures unifiables. Les sommets du graphe lexical s'apparentent alors à des objets disposant d'un certain nombre d'Attributs, disponibles dans leur description lexicographique. Ils entretiennent des Relations, représentées par les arcs. Une réflexion est menée sur la nature des différents éléments composant le réseau et sur les différents rapports qu'ils entretiennent entre eux. Elle est réalisée en prenant en compte l'évolution de la ressource sur une période de trente mois. Elle est accompagnée d'une analyse topologique, qui met en avant des propriétés proches de celles des graphes petit monde. Deux séries d'expériences exploratoires sont ensuite réalisées. La première d'entre elles permet de conforter l'idée selon laquelle la formalisation en œuvre dans la ressource permet de détecter automatiquement des analogies conformes à l'intuition des locuteurs. Elle met en avant la possibilité de réaliser différents types d'exploration par raisonnement analogique, en fonction des points d'entrée et des éléments d'informations comparés. Elle montre également l'apport de telles explorations en terme de vérification de la cohérence du réseau et d'émergence de règles lexicales. La seconde série d'expériences se concentre autour de la notion de configurations de dérivations lexicales. Elle montre comment le regroupement de sous-graphes analogues met en avant l'existence de connexions lexicales récurrentes à travers la ressource. L'état d'avancement de la ressource exploitée ne permet pas d'obtenir des règles et des modèles aboutis. Les résultats obtenus sont toutefois encourageants. Les observations réalisées nous amènent à considérer l'analogie comme un guide permettant de s'assurer de la bonne qualité de la représentation du lexique proposée par une ressource. Elle permet également d'acquérir automatiquement des connaissances sur son organisation. De telles connaissances permettent d'identifier des phénomènes linguistiques et d'instrumenter l'activité lexicographique.



**Mots clés :** lexicographie - analogie - raisonnement analogique - système lexical - graphe petit monde



## Abstract

By using theorical and practical advances in computer sciences and linguistics, contemporary lexicography have freed itself from the linear organisation of print publications. It had sought to define new descriptive models and provide henceforth the community with formal and consistent ressources. Such ressources offer many opportunities for natural language processing tasks. This PhD thesis focuses on lexical systems model, which is proposed by Explanatory Combinatorial Lexicography. Specifically, it takes an interest in the the French Lexical Network, presently under development. As lexical system, this ressource is a monolingual graph of lexical units connected by a rich set of lexical relations mainly based on syntactico-semantic Meaning-Text lexical functions. The PhD thesis lays the groundwork for an exploration of this ressource by analogical reasonning. It begins with a selective overview of formalisation and computerisation of analogy in natural language processing. It thus defines the principle of exploration of lexical graph implemented as a grouping of unifiable structures. The nodes in the graph are similar to objects, which have some attributes available in their lexicographic descriptions and edges represent relations. A reflection is conducted on the nature of the various constituents of the graph and the relations between them. The evolution of the resource over thirty months is taken into account during the carrying out of this study. It is accompanied by a topological analysis, which highlights properties close to those of small word networks. Then two sets of exploratory experiments are conducted. The first one confirms the belief that the resource formalisation makes it possible to detect automatically analogies consistent with lexicographers'intuition. It highlights the possibility of several kind of analogical explorations, according to entry points and compared pieces of information. It also reveals that this exploration enable us to check the consistency of the resource and to allow lexical rules to emerge. The second set of experiments is focused around the concept of lexical derivation configurations. It shows how grouping of analogous subgraphs reveals recurrent lexical connections through the resource. The progress status of the resource doesn't enable us to obtain successfully completed rules and models. The results obtained are nontheless encouraging. Analogy can already be considered as a guide to ensure the quality of lexical resources. It also allows for the acquisition of knowledge about its organisation. Such knowledge can be used to identify linguitic phenomena and to design instruments to support lexicographic activity.



**Keywords :** lexicography - analogy - analogical reasoning - lexical system - small word network