Making Use of Existing Lexical Resources to Build a Verbnet like Classification of French Verbs

Ingrid Falk

▶ **To cite this version:**

Ingrid Falk. Making Use of Existing Lexical Resources to Build a Verbnet like Classification of French Verbs. Computation and Language [cs.CL]. Université Nancy II, 2012. English. NNT : . tel-00714737

HAL Id: tel-00714737

https://theses.hal.science/tel-00714737v1

Submitted on 5 Jul 2012

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

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

UNIVERSITÉ DE LORRAINE

UFR mathématiques et informatique **École doctorale IAEM Lorraine**

Département de formation doctorale en informatique

Making Use of Existing Lexical Resources to Build a Verbnet like Classification of French Verbs

THÈSE

présentée et soutenue publiquement le 13 juin 2012

pour l'obtention du

Doctorat de l'Université de Lorraine

(spécialité informatique)

par

Ingrid Falk

Composition du jury

Rapporteurs :	Sabine Schulte im Walde	Université de Stuttgart, Allemagne
	Lonneke van der Plas	Université de Stuttgart, Allemagne
Examinateurs :	Claire Gardent (Directrice)	CNRS, Nancy, France
	Samuel Cruz-Lara (Co-Directeur)	Université de Lorraine, Nancy, France
	Anne Boyer	Université de Lorraine, Nancy, France
	Thierry Poibeau	CNRS, Paris, France
Invité :	Jean-Charles Lamirel	Université de Strasbourg, France

Laboratoire Lorrain de Recherche en Informatique et ses Applications — UMR 7503

Loria

Mis en page avec la classe thloria.

Acknowledgements

I very much enjoyed working on this PhD thesis and there are so many people and circumstances that contributed to turning this into a scientifically and personally enriching experience that it can not be possible to give credit to all.

First and foremost I would like to thank my advisor, Claire Gardent. Without her agreeing to take me as her advisee, in spite of my uncommon personal and professional situation and background, this thesis would not have been possible in the first place. I am very grateful for the many things Claire taught me by being herself a passionate researcher and a good friend.

I am also indebted to Samuel Cruz-Lara for supporting me on the SEMbySEM project and giving me the opportunity to experience an industrial research project.

Jean-Charles Lamirel also contributed significantly to this work. He not only introduced Claire and me to the IGNGF clustering method, but also performed countless computations and shared his intuition and insights for visualising and interpreting the results. I am very grateful for his dedicated participation and involvement.

Many thanks also go to the reviewers of this thesis, Sabine Schulte im Walde and Lonneke van der Plas, who so thoroughly reviewed my thesis and provided me with very interesting and useful feedback. I am indebted to all the members of the jury for showing interest in this work and for turning its defence into an interesting and stimulating discussion.

At the Loria Lab and in particular in the Talaris/Synalp team I experienced an international and stimulating environment which made the work at this thesis agreeable and exciting and I am indebted to all the people who made this happen, in particular Lina Rojas Barahona and Alexandre Denis with whom I shared an office for many months. Special thanks go to Emmanuel Didiot, who agreed to proofread the French summary of this thesis and who turned its awkward French into a more fluent and easy to read version.

Finally, I would like to thank my husband Jens Gustedt for his unconditional support and encouragement and I am also very grateful to our daughters Lea and Kari for enthusiastically approving of this endeavour.

Résumé

Mots-clés: Classification de verbes français, classification verbale syntactico-sémantique, Verbnet, Analyse Formelle des Conceptes, AFC, algorithme incrémental de gaz neuronal croissant avec maximisation de vraisemblance, IGNGF

Des classifications verbales associant classes de verbes avec des propriétés syntaxiques et sémantiques communes aux membres d'une classe se sont montrées utiles aussi bien dans la recherche linguistique que dans le traitement automatique des langues. Cette thèse a pour objectif de présenter des approches pour l'acquisition automatique de classes verbales pour le Français palliant ainsi partiellement le manque de ce type de ressources pour le Français. Par rapport aux classes générées, dans la plupart des approches existantes, les classes de verbes produites ne sont pas associées avec une caractérisation explicite des propriétés syntaxiques et sémantiques partagées par les membres des classes. Notre approche permet non seulement de créer des classes de verbes mais aussi d'associer ces classes avec les cadres de sous-catégorisations et les grilles thématiques partagés par les membres d'une classe.

Nous utilisons deux méthodes de classification pour acquérir des classes verbales. La première est une méthode symbolique appelée *Analyse Formelle de Conceptes* (FCA - Formal Concept Analysis). La deuxième exploite un algorithme de gaz neuronal croissant basé sur l'étiquetage des clusters par maximisation de vraisemblance (IGNGF - Incremental Growing Neural Gas with Feature maximisation). Pour la création des classes verbales, nous appliquons ces deux méthodes aux même ressources Françaises et Anglaises. Celle-ci sont constituées d'une part d'un lexique syntaxique pour les verbes du Français, issue de la fusion de trois ressources pour le Français existantes. D'autre part elles sont obtenues par traduction automatique en Français des classes du Verbnet anglais. Les classes verbales produites sont associées à des informations syntaxiques et sémantiques explicites sous forme de cadres de sous-catégorisations et grilles thématiques.

Les classifications produites sont évaluées dans un premier temps en tant que groupements de verbes par une comparaison à une référence (proposé par [Sun *et al.*, 2010]). Deuxièmement, les associations aux cadres syntaxiques et aux grilles thématiques sont évaluée d'une part d'une façon intrinsèque par une comparaison à une annotation manuelle en rôles thématiques. D'autre part nous effectuons une évaluation extrinsèque en utilisant les classes verbales dans une tâche d'annotation en rôles thématiques simplifiée.

Ces évaluations montrent que les classifications obtenues par les deux méthodes sont pertinentes tant par rapport aux groupement de verbes produits qu'aux associations de ces verbes avec des cadres de sous-catégorisation et des grilles thématiques. Elles présentent néanmoins des caractéristiques complémentaires. Tandis que les classes produites par FCA se sont révélées plus performantes par rapport aux associations ⟨verbe, cadre syntaxique⟩ et ⟨verbe, grille thématique⟩, les classes générées par IGNGF correspondent mieux à la classi-

fication de référence et se sont montrées plus efficaces à l'attribution de rôles thématiques.

Abstract

Keywords: Verb classification for French, syntactic-semantic verb classification, Verbnet, Formal Concept Analysis, FCA, Incremental Growing Neural Gas with Feature Maximisation, IGNGF

Classifications which group together verbs and a set of shared syntactic and semantic properties have proven useful both in linguistics and in Natural Language Processing tasks. However, for French this type of classifications is not available in a format suitable for automated processing. In addition, most existing approaches for automatically acquiring verb classes fail to associate the verb classes produced with an explicit characterisation of the syntactic and semantic properties shared by the class members. Here we propose a novel approach to verb clustering which addresses these shortcomings. We classify French verbs using two clustering methods, a symbolic method called Formal Concept Analysis (FCA) and a probabilistic neural clustering method called Incremental Growing Neural Gas with Feature Maximisation (IGNGF). The obtained classes group together verbs, subcategorisation frames and thematic grids. We apply this approach to French data consisting of roughly 4000 verbs and 350 subcategorisation frames, and evaluate both the clusters obtained (i.e., verb classes) and the features labeling each cluster (i.e., syntactic frames and thematic grids). The results suggest that both classification methods can be used to bootstrap a Verbnet style classification for French such that the verb classes it contains (i) are reasonably clean and (ii) associate verbs with partial information about subcategorisation frames and thematic grids. The obtained classifications are complementary. While the FCA classification better represents verb polysemy (better F-measure and recall compared to reference data) the IGNGF classification performed better with respect to the produced verb classes and when used in a task based evaluation.

Contents

F Acquisition de classes verbales pour le français

This chapter presents a summary of the thesis, in French.
Ce chapitre présente un résumé en français de la thèse.

Sommaire

Un but du traitement automatique des langues est de concevoir des outils permettant à des automates de comprendre le sens d'un texte en langue naturelle. Les éléments linguistiques donnant accès au sens d'un texte sont surtout des prédicats, en général des verbes, qui se combinent syntaxiquement avec d'autres mots pour représenter des événements, les participants à ces événements et leurs rôles sémantiques. Par conséquent, des connaissances sur les verbes et leur comportement syntaxique et sémantique sont primordiales. Ce type d'information est souvent représenté sous forme de lexiques de verbes.

> Ce travail consiste principalement en la constitution et l'évaluation d'un tel lexique de verbes pour le Français.

Si un nombre important d'approches pour la constitution automatique de lexiques de verbes a été proposé, actuellement il n'existe pas de consensus sur la meilleure façon de créer de telles ressources adaptées au traitement automatique des langues. Pourtant, ces représentations, regroupant des verbes par rapport à leurs propriétés syntaxiques et sémantiques et explicitant clairement ces propriétés se sont souvent montrées utile aussi bien dans la recherche linguistique que celle du traitement automatique de la langue.

> Cette thèse présente *une approche pour l'acquisition automatique d'une classification syntactico-sémantique des verbes du Français.* Nous construisons des classes de verbes dont le comportement syntaxique et sémantique est explicité en associant chaque classe avec un ensemble de cadres de sous-catégorisation d'une part (syntaxe) et un ensemble de rôles thématiques d'autre part (sémantique).

Pour l'anglais, ce type de classification a été réalisé avec *VerbNet* ([Schuler, 2006]). *VerbNet* est un lexique verbal électronique qui reprend la classification de Beth Levin et l'étend systématiquement en assurant la cohérence sémantique et syntaxique de ses classes. Toujours pour l'anglais, il existe plusieurs ressources proposant des classes de verbes dans un format adapté au traitement automatique : Framenet [Baker *et al.*, 1998] et dans une moindre mesure Wordnet [Fellbaum, 1998].

> Pour ces travaux nous visons une classification proche de Verbnet, d'une part en raison de ses fondements théoriques découlant de l'ancrage de Verbnet dans les travaux de Beth Levin, et d'autre part pour sa large couverture, qui rend ce lexique particulièrement utile pour les applications en traitement automatique des langues. En outre, en nous basant sur Verbnet, nous souhaitons mettre à profit la recherche considérable qui a menée à sa construction afin d'obtenir une ressource semblable pour une autre langue, le Français. Enfin, l'un de nos objectifs à plus long terme est d'utiliser la ressource créée dans une tâche d'annotation en rôles sémantiques pour le Français. En effet, Verbnet semble fournir une représentation de l'information propice à l'utilisation pour l'annotation sémantique en Français : d'une part l'ensemble de rôles thématiques Verbnet est moins dépendant de la langue, et d'autre part il s'est montré adapté dans des tâches similaires pour l'Anglais.

Il existe relativement peu d'études concernant la constitution de classifications de type Levin pour d'autres langues que l'Anglais ([Sun *et al.*, 2010; Brew and Schulte im Walde, 2002; Schulte im Walde, 2003; 2006; Oishi and Matsumoto, 1997; Dang *et al.*, 1998; Merlo *et al.*, 2002]). La plupart d'entre elles se concentrent sur la création des classes à l'aide de traits extraits d'un corpus.

> Puisque notre objectif est d'obtenir une classification couvrants les verbes principaux du Français, nous avons choisi d'utiliser des traits extraits de ressources lexicales validées manuellement, plutôt que des données distributionnelles.

La majorité des travaux dans ce domaine se contente de constituer des ensembles de verbes cohérents d'un point de vue syntaxique et sémantique. Cependant, les traits justifiant cette cohérence restent implicites : ils déterminent le groupement de verbes similaires en classes, mais les traits ne sont pas utilisés pour étiqueter les classes. Par conséquent, l'apport de chaque trait à la constitution d'une classe n'apparaît pas clairement.

> Dans l'approche présentée ici, les classes verbales sont explicitement associées à des cadres syntaxiques et des rôles thématiques. Les associations avec des cadres syntaxiques découlent directement de la méthode de classification. Pour l'attribution des rôles thématiques nous nous appuyons sur l'hypothèse que les composantes sémantiques des classes verbales du Verbnet ne sont pas spécifiques à l'anglais ([Jackendoff, 1990]) et asso-

cions les groupes de verbes avec les rôles thématiques des classes traduites de l'anglais.

En prenant comme point de départ des ressources lexicales Françaises et Anglaises, nous constituons des classes verbales en utilisant deux méthodes de classification. La première est une technique symbolique : l'*Analyse Formelle de Concepts* (FCA). La seconde emploie un algorithme incrémental de "gaz neuronal croissant" basé sur l'étiquetage de clusters par maximisation de vraisemblance (IGNGF - Incremental Growing Neural Gas with Feature maximisation). Les classifications obtenues par ces deux méthodes sont par la suite analysées et évaluées suivant plusieurs schémas. Elles sont d'abord évaluées selon les groupements de verbes produits, en les comparant à une référence proposée dans la littérature ([Sun *et al.*, 2010]). Cette référence est ensuite utilisée pour une évaluation sémantique : nous vérifions dans quelle mesure nos classifications produisent les mêmes associations ⟨verbe, rôles thématiques⟩ que la référence.

Afin d'évaluer la cohérence syntaxique de nos classes et d'estimer leur contribution possible à l'étiquetage en rôles thématiques, nous étudions les associations ⟨verbe, cadre syntaxique, rôles thématiques⟩ A cette fin nous avons créé une deuxième référence (appelée SRL gold) en annotatant manuellement des verbes et leurs arguments syntaxiques dans un corpus arboré français ([Abeille *et al.*, 2003; Kupść and Abeillé, 2008]). Cette référence permet de comparer les associations ⟨verbe, cadre syntaxique, rôles thématiques⟩ générées par nos classifications aux annotations du corpus. Enfin, nous utilisons la référence SRL gold pour effectuer une évaluation axée sur une tâche. Nous examinons la capacité de nos classifications à aider à l'attribution de rôles thématiques aux verbes et à leurs arguments, en comparant les étiquettes attribués par le système à celles de la référence.

L'approche pour l'acquisition automatique de classes verbales à la Beth Levin, que nous proposons ici, pourrait être étendue à d'autres langues à condition de disposer des ressources nécessaires (lexique syntaxique et dictionnaire de traduction de l'anglais). Elle pourrait également être appliquée sur à des données extraites d'un corpus, ce qui la rendrait totalement non-supervisée et applicable à toute langue pour laquelle un analyseur syntaxique serait disponible.

La thèse tout comme le résumé, est structurée de la manière suivante. Dans la première section nous présentons les ressources et méthodes utilisées dans ce travail pour l'acquisition de classes verbales. Nous y décrivons les ressources lexicales suivantes : un lexique syntaxique pour le français, le lexique verbal Verbnet pour l'anglais

et trois dictionnaires français-anglais. Nous introduisons ensuite les techniques de classification que nous avons utilisées : l'analyse formelle de concepts (FCA) et le clustering incrémental de gaz neuronal croissant basé sur l'étiquetage des clusters par maximisation de vraisemblance (IGNGF). Nous montrons aussi comment ces méthodes sont appliquées à nos ressources pour créer des classes verbales associées à des ensembles de cadres syntaxiques et rôles thématiques. Finalement nous décrivons les ensembles de données utilisées pour l'évaluation, composées d'une part de la classification proposée dans [Sun *et al.*, 2010] (appelée ici V-gold) et d'autre part d'une référence appelée ici SRL-gold, qui a été obtenue par annotation manuelle en rôles thématiques des verbes et de leurs arguments dans un corpus arborée français.

La section suivante (Section F.2) est dédiée à l'évaluation sémantique des classes obtenues : Nous évaluons la cohérence des groupes de verbes produits en les comparant à la référence V-gold de deux manières : d'abord d'une manière globale basées sur des mesures utilisées en clustering, puis dans un deuxième temps par rapport aux associations des verbes avec rôles thématiques.

L'évaluation présentée en Section F.3 est axée sur l'association des classes verbales avec des ensembles de propriétés syntaxiques (cadres de sous-catégorisation) et sémantiques (rôles thématiques). Nous vérifions dans quelle mesure les associations des verbes avec des cadres syntaxiques et rôles thématiques, présentes dans la référence SRL-gold, sont générées par nos classifications. Également par une comparaison à la référence SRL-gold, nous examinons la capacité des classes verbales de prédire des associations des verbes et leurs arguments syntaxiques avec des rôles thématiques Verbnet.

La conclusion de ce mémoire résume nos contributions et mentionne quelques perspectives que nous inspirent ces travaux.

1 Ressources et méthodes

1.1 Ressources lexicales et attributs.

1.1.1 Ressources syntaxiques.

Les traits syntaxiques et sémantiques que nous utilisons pour l'acquisition des classes verbales sont extraits de ressources lexicales existantes. Les traits syntaxiques sont principalement des cadre de sous-catégorisation provenant des lexiques français Dicovalence, TreeLex et LADL. Comme les lexiques Dicovalence et LADL contiennent une information plus riche qui n'est pas toujours représentée dans les cadres syntaxiques, nous extrayons de ces lexiques des traits syntaxiques et sémantiques supplémentaires,

susceptibles d'aider à l'identification des rôles thématiques caractérisant un certain usage du verbe. Par la suite nous décrivons brièvement ces lexiques en précisant les traits syntaxiques et sémantique extraits.

Dicovalence ([Mertens, 2010]) est une ressource informatique qui répertorie les cadres de valence de plus de 3700 verbes simples du français. Il se présente comme une liste d'entrées correspondant chacune à un emploi d'un lemme verbal, associant a ce lemme un cadre de sous-catégorisation et différents pronoms qui peuvent être une réalisation d'un des arguments présent dans le cadre. Les traits syntaxiques que nous extrayons de ces entrées sont d'abord les cadres de sous-catégorisation. En plus de ces cadres, les réalisations pronominales indiquées dans Dicovalence, permettent de déduire certains traits syntaxiques et sémantiques que nous utilisons également pour l'acquisition des classes. Les traits syntaxiques sont les suivants :

Sym: indication de(s) arguments symétriques (*L'un accuse l'autre des pires mensonges*),

ArgNbr: plus de quatre arguments syntaxiques,

Event: présence d'un argument phrastique,

Pred: présence d'un complément prédicatif,

Theme: présence complément d'objet optionnel ou passivation avec "se" (d'après [Randall, 2010], p. 95 et p. 120 resp., ce comportement syntaxique indique un rôle thématique Theme).

Les traits sémantiques sont :

Loc: présence d'un argument locatif ou dé-locatif,

Nhum: indication d'un argument non-humain,

Asset: présence d'un argument de quantité,

Plural: indication d'un argument nécessairement au pluriel.

TreeLex [Kupść and Abeillé, 2008] est un lexique de sous-catégorisation développé automatiquement à partir d'un corpus annoté en dépendances syntaxiques (treebank). Il comprend environs 2000 verbes du français contemporain (types) avec leurs cadres de sous-catégorisation, 180 cadres syntaxiques distinct et approximativement 2.09 cadres par verbe.

LADL. Le lexique-grammaire de Maurice Gross et sa forme électronique dérivée, les tables du LADL, donnent une description systématique des foncteurs syntaxiques du français. Il est partiellement disponible en version électronique et contient des informations de sous-catégorisation qui sont à la fois détaillées et extensives. Le lexique-grammaire est organisé en un ensemble de tables, chaque table regroupant les usages des mots prédicatifs qui partagent les propriétés dites définitoires de la table. En particulier, toutes les entrées d'une table ont en commun un (parfois deux) cadre(s) de sous- catégorisation de base. C'est cette informations que nous utilisons comme traits syntaxiques. En dessus des cadres syntaxiques, nous déduisons des traits syntaxiques et sémantiques à partir de l'appartenance des verbes à certaines tables. Les traits syntaxiques sont inférés comme suit :

Sym: présence d'arguments symétriques, indiqués par l'appartenance aux Tables 36S, 36SL et 35S,

ArgNbr: plus de quatre arguments syntaxiques : Tables 18 et 38L

Pred: la Table 39 regroupe des verbes à constructions à attributs (*On a nommé Paul président*),

Event: présences d'arguments phrastiques : Tables 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16

Nous extrayons les traits sémantiques suivants :

Loc: rôle locatif ou dé-locatif : Tables 2, 3, 35L, 37E, 38LS, 38LH, 38LD, 38L, 35ST, 34L0, 37M4, 37M5, 37M6 and 38L1

Nhum: rôle non-humain : Tables 32C, 32A et 32CV

Asset: rôle indiquant une quantité : Table 32NM

Plural: rôle nécessairement au pluriel : Tables 32PL, 38PL, 36S, 35S et 36SL

Lexique syntaxique fusionné. L'information de sous-catégorisation présente dans les trois lexiques a été uniformisé et combiné résultant en un lexique de 20443 entrées ⟨verbe, cadre⟩ (5918 verbes et 345 cadres de sous-catégorisation). Le Tableau 1 montre les entrées correspondant au verbe *expédier*. Ce tableau montre aussi le format des cadres syntaxiques : pour chaque fonction sous-catégorisé (SUJ - sujet, OBJ - objet direct, DEOBJ - objet indirect en *de*, AOBJ - objet indirect en *à*, POBJ - complément avec une préposition autre que *à* ou *de* et ATB - attribut

Verb: *expédier*	
SCF	**Source info**
SUJ:NP,DUMMY:REFL	DV:41640,41650
SUJ:NP,OBJ:NP	DV:41640,41650;TL
SUJ:NP,OBJ:NP,AOBJ:PP	TL
SUJ:NP,OBJ:NP,POBJ:PP	DV:41640
SUJ:NP,OBJ:NP,POBJ:PP,POBJ:PP	LA:38L
SUJ:NP,OBJ:NP,POBJ:VPinf	LA:3
SUJ:NP,POBJ:PP,DUMMY:REFL	DV:41640

Tableau 1: Exemple d'entrées du lexique de sous-catégorisation fusionné pour le verbe *expédier*. La troisième colonne montre la ressource originale: DV (Dicovalence), LA (LADL tables) ou TL (Treelex).

du sujet ou de l'objet) le cadre spécifie la catégorie syntaxique (NP - syntagme nominal, PP - syntagme prépositionel, VPinf - syntagme verbal, groupe infinitif, Ssub - proposition). Cette ressource est librement disponible à l'adresse suivante: `http://talc.loria.fr/tl_dv2_ladl-a-subcategorisation.html`.

Les traits. Les attributs (syntaxiques) utilisés pour l'acquisition des classes sont d'une part les cadres syntaxiques extraits du lexique syntaxique fusionné : On considère qu'un verbe a un cadre syntaxique si il est associé avec ce cadre dans le lexique syntaxique. D'autre part, nous utilisons les traits syntaxiques *Sym, ArgNbr, Pred, Event, Theme* et sémantiques *Loc, Nhum, Asset, Plural* extraits des lexiques Dicovalence et LADL. Ces traits sont attribués aux verbes comme expliqué précédemment.

1.1.2 Ressources sémantiques

La ressource que nous utilisons pour l'étiquetage sémantique de groupes de verbes français est le Verbnet anglais. Nous alignons ces groupes de verbes français avec les classes Verbnet traduites et leur attribuons les rôles thématiques de la classe Verbnet associée. Les classes traduites sont utilisées de deux façons. Premièrement elles sont alignées avec des groupes de verbes afin de pouvoir associer ces derniers avec des rôles thématiques. Deuxièmement, nous en dérivons des traits sémantiques qui contribuent aux processus de classification.

Verbnet est un lexique verbal électronique qui reprend la classification de Beth Levin et l'étend systématiquement en assurant la cohérence sémantique et syntaxique de ses classes. Chaque classe Verbnet est associé avec un ensemble de cadres de sous-catégorisation représentant l'usage syntaxique des membres et avec un ensemble de rôles thématiques désignant les participants aux évènements évoqués par les verbes membres. La Figure 1 donne un aperçu de la classe Verbnet *amuse-31.1*

verbes (242): abash, affect, afflict, amuse, annoy, ...

rôles thématiques: EXPERIENCER [+animate]
CAUSE

cadres syntaxiques (6):

NP V NP	CAUSE V EXPERIENCER
NP V ADV-Middle	EXPERIENCER V Adv
NP V	CAUSE V
NP V NP PP.oblique	CAUSE V EXPERIENCER *with* OBLIQUE
NP.cause V NP	CAUSE('s) OBLIQUE V EXPERIENCER
NP V NP ADJ	CAUSE V EXPERIENCER Adj
...	

Figure 1: La classe Verbnet *amuse-31.1* - exemple simplifié.

avec ses cadres syntaxiques et rôles thématiques associés. Cet exemple montre que les classes Verbnet explicitent la réalisation syntaxique des rôles thématiques : Dans la figure il est spécifié que dans le cadre syntaxique NP V NP, le sujet est la réalisation syntaxique du rôle thématique *Experiencer* alors que le rôle *Cause* est réalisé syntaxiquement par l'objet direct.

Pour traduire les verbes des classes Verbnet anglaises nous utilisons trois dictionnaires Français-Anglais: Sci-Fran-Euradic, Google dictionary et Dicovalence. Sci-Fran-Euradic est un dictionnaire bilingue créé par des linguistes et distribué par ELDA (`http://catalog.elra.info/product_info.php?products_id=666`). Il contient 40 111 paires de verbes Français-Anglais. De google dictionary nous avons extraits 13 824 paires de verbes et 11 351 de Dicovalence. Cette manière de traduire des classes pose néanmoins des problèmes dus principalement à la polysémie verbale[1]. Pour pallier ce problème nous avons appliqué deux méthodes pour traduire les classes Verbnet. La première est basée sur la fréquence des traductions et la deuxième s'appuie sur la technique d'apprentissage supervisée à machine à vecteurs de support (SVM - support vector machines). Nous allons maintenant donner un aperçu de l'application de ces méthodes à notre tâche.

En ce qui concerne la méthode basée sur la **fréquence des traductions**, appelée **median** par la suite, les classes Françaises sont constituées comme suit. Pour une classe Verbnet anglaise nous traduisons chaque verbe membre en Français et classons les traductions obtenues par nombre décroissant de traductions anglaises membre de la classe Verbnet de départ. Nous ne gardons dans la classe Verbnet française que les verbes de la moitié supérieure de ce classement.

Pour ce qui est de la traduction des classes par **SVM** le problème de la traduction est reformulé en une tâche de classification binaire qui est de décider si un

[1]La polysémie moyenne attesté dans le WordNet (anglais) est de 2.17 pour verbes, 1.4 pour adjectives, 1.25 pour adverbes et de 1.24 pour noms.

verbe français est dans une classe Verbnet traduite ou non. Dans ce contexte, les objets à classifier sont des paires de verbes français et de classes Verbnet anglaises. Pour créer un classifieur SVM il est nécessaire d'une part de constituer un ensemble d'apprentissage, d'autre part d'attribuer des traits aux objets à classifier.

L'ensemble d'apprentissage que nous utilisons désigne pour chaque pair ⟨verbe français, classe Verbnet anglaise⟩ si on peut considérer (ou non) que le verbe a un sens correspondant à la classe Verbnet anglaise. Le classifieur obtenue par l'apprentissage sur cet ensemble rend, pour chaque pair une estimation, interprétée comme la probabilité du verbe français d'être un membre de la classe Verbnet anglaise. Nous sélectionnons les paires les plus pertinentes par rapport à ces probabilités et constituons les classes traduites en attribuant les verbes des paires sélectionnées aux classes Verbnet correspondantes.

Les traits que nous attribuons aux objets à classifier sont numériques et similaires aux scores utilisés dans [Mouton, 2010] : ils sont dérivés par exemple du nombre total des traductions d'un verbe anglais ou français, la taille des classes Verbnet, le nombre de classes dont un verbe est membre, etc.

Basé sur ces éléments nous utilisons libsvm[2], une implémentation d'un classifieur SVM, pour générer des probabilités estimées pour chaque paire ⟨verbe français, classe Verbnet anglaise⟩. À partir de ces estimations nous produisons trois ensembles de classes traduites, en fonction des critères de sélection des paires les plus pertinents. Pour le premier, nommé **svm** nous sélectionnons les 6000 paires avec la probabilité la plus élevée[3]. Pour les classes **svm-median** nous sélectionnons pour chaque verbe la paire ⟨verbe, classe⟩ avec la meilleure probabilité. Enfin, pour l'ensemble **svm-median** nous choisissons pour chaque verbe, les paires dont la probabilité est en dessus du médian (par rapport à ce verbe).

Comparé à Verbnet, l'ensemble le plus proche par rapport à la distribution des verbes en classes est **svm**. En Section F.1.3 nous allons utiliser chacun de ces ensembles pour associer des rôles thématiques aux classes verbales. Le résultat de ces expériences indique également que les classes **svm** sont les mieux adaptées à notre tâche.

Les traits. Les classes Verbnet traduites sont utilisées comme traits sémantiques de la manière suivante. Premièrement elles sont utilisées pour caractériser des groupes de verbes français. Pour cela, des groupes de verbes français sont alignés à des classes traduites sur la base des verbes membre communs. Un groupe de verbe

[2] `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`

[3] Nous avons choisit ce nombre en raison de la quantité de paires ⟨verbe, classe⟩ dans Verbnet, qui est de 5726.

français sera associé à l'ensemble des rôles thématiques de la classe Verbnet la plus proche. Deuxièmement, les classes traduites sont utilisés comme attributs dans le processus de classification : on considère qu'un verbe "a" une classe Verbnet si il est membre de la classe traduite correspondante.

Les différents ensembles d'attributs décris ici sont combinés de plusieurs façons résultant en huit ensembles de traits que nous avons utilisé pour constituer les classes verbales :

scf cadres de sous-catégorisation uniquement

scf-synt cadres de sous-catégorisation et traits syntaxiques

scf-sem cadres de sous-catégorisation et traits sémantiques

scf-synt-sem cadres de sous-catégorisation et traits syntaxiques et sémantiques

grid-scf cadres de sous-catégorisation et traits dérivés des classes Verbnet traduites

grid-scf-synt cadres de sous-catégorisation, traits syntaxiques et classes Verbnet traduites

grid-scf-sem cadres de sous-catégorisation, traits syntaxiques et classes Verbnet traduites

grid-scf-synt-sem cadres de sous-catégorisation, traits syntaxiques et sémantiques et classes Verbnet traduites.

1.2 Jeux de données

Nous décrivons maintenant les jeux de données utilisés dans nos expériences et les données de référence à la base de nos évaluations. Ces données sont présenté ensemble car, pour rendre une comparaison possible, nous nous sommes assurés de la compatibilité des données utilisées pour le clustering et celles de référence. Nous introduisons d'abord les données de référence. Pour le Français, il existe une seule ressource susceptible d'être utilisée dans l'évaluation de classes verbales. Cette ressource, nommée V-gold dans la suite, a été proposée dans [Sun *et al.*, 2010] et regroupe 160 verbes français dans des classes verbales à la Beth Levin. Nous identifions ces classes verbales avec des classes Verbnet et obtenons de cette façon des associations ⟨verbe, rôles thématiques⟩ qui peuvent être comparées avec les associations induites par nos classifications. Cette ressource ne permet cependant pas une évaluation des associations des classes verbales avec des cadres syntaxiques. Pour pouvoir évaluer les associations ⟨verbe, cadre⟩ générées par nos classifications nous avons créer une deuxième

Data set	# verbes	# scfs	# rôles thématiques	# classes traduites
vn_restricted	2091	238	13	70
vn_all	4260	303	16	77

Tableau 2: Dimensions des ensembles de données utilisés dans les expériences d'acquisition de classes verbales.

ressource de référence, appelée SRL-gold, où nous avons annoté en rôles thématiques Verbnet les instances ⟨verbe, argument syntaxique⟩ d'un corpus arboré français.

Pour positionner nos travaux par rapport à la référence V-gold, nous nous sommes limité aux classes Verbnet présentes dans cette ressource. Comme les rôles thématiques Verbnet, choisis sur des critères linguistiques, ne sont pas toujours adaptés à des méthodes automatisées, nous les avons regroupés de telle façon que chaque classe Verbnet soit caractérisée par un ensemble unique et bien défini de rôles thématiques. Les rôles thématiques retenues pour cette série d'expérience sont les suivants : AgExp, AgentSym, Theme, PredAtt, ThemeSym, Patient, PatientSym, Start, End, Location, Beneficiary, Cause et Instrument. Chaque classe Verbnet est identifiée à l'ensemble de ses rôles thématiques et les classes avec les mêmes rôles thématiques sont regroupées. Ensuite, ces classes sont traduites et l'ensemble de rôles thématiques détermine l'étiquetage sémantique des classes verbales que nous construisons.

Dans les expériences où nous comparons nos résultats avec la référence SRL-gold, nous ciblons une classification à large échelle et utilisons toutes les classes Verbnet. De ce fait, l'ensemble de rôles thématiques utilisé et pour étiqueter les classes verbales et pour annoter les instances de SRL gold est plus étendu : Agent, AgentSym, Experiencer, Patient, PatientSym, Theme, ThemeSym, Topic, PredAtt, Start, End, Location, Beneficiary, Cause, Extent, Instrument.

Nous avons effectué nos expériences sur deux jeux de données. Le premier (**vn_restricted**) est basé sur l'ensemble de classes Verbnet et les rôles thématiques issue de la référence V-gold. Le deuxième (**vn_all**) est constitué en utilisant toutes les verbes du lexique syntaxique fusionné et toutes les classes Verbnet. Les classes Verbnet anglaises sont regroupées en fonction de l'ensemble des rôles thématiques correspondants et ensuite sont traduites suivant les méthodes présentées plus haut. Les expériences basées sur le jeu de donnée **vn_restricted** ont pour but, entre autre, de déterminer la méthode de traduction la plus efficace pour notre tâche, qui est la méthode *svm*. Ensuite, pour les expériences **vn_all** les classes Verbnet sont traduites en utilisant cette méthode. Le Tableau 2 montre les dimensions des jeux de données résultants.

1.3 Méthodes de classification

Nous avons appliqué deux techniques de clustering/classification pour produire des classes verbales à la Verbnet. La première est une méthode symbolique, l'analyse formelle de concepts (FCA - Formal Concept Analysis), la deuxième est une méthode probabiliste s'appuyant sur un algorithme incrémental de gaz neuronal croissant basé sur l'étiquetage des clusters par maximisation de vraisemblance (IGNGF - Incremental Growing Neural Gas with Feature maximisation). La méthode FCA regroupe simultanément les verbes en fonction des cadres syntaxiques qu'ils acceptent et les cadres syntaxiques en fonction des verbes pour lesquels ils sont acceptés, tandis que IGNGF regroupe les verbes selon les attributs (traits) extraits du lexique. En conséquence, FCA produit naturellement des classes verbales associées avec des ensembles de cadres de sous-catégorisation, qui, pour obtenir des classes à la Verbnet, doivent être associées à un ensemble de rôles thématiques. Par contre, IGNGF produit des groupes de verbes associés aux traits (cadres syntaxiques) les plus pertinents pour la constitution du cluster. Pour arriver aux classes syntactico-sémantiques ciblées, il est donc nécessaire d'associer ces clusters avec des cadres de sous-catégorisation et des grilles thématiques.

Par la suite nous présentons d'abord les fondements théoriques de ces deux méthodes avant de montrer comment elles sont appliquées à l'acquisition des classes Verbnet. En particulier, pour FCA nous décrivons comment cette technique est utilisée pour créer des classes de verbes associées à des ensembles de cadres de sous-catégorisation et ensuite comment à ces classes sont attribués des rôles thématiques. Pour IGNGF nous montrons d'abord comment les verbes sont regroupés en fonction des traits extraits des ressources lexicales décrites en Section F.1.1 et ensuite comment les clusters obtenus sont associés à des cadres de sous-catégorisation et rôles thématiques.

1.3.1 FCA.

L'analyse formelle de concepts (Formal Concept Analysis - FCA, [Barbut and Monjardet, 1970; Ganter and Wille, 1999]) est une technique de classification qui permet de créer, en partant d'un *contexte formel*, un treillis de concepts, où les concepts sont constitués d'un ensemble d'objets et de l'ensemble d'attributs que ces objets partagent. Dans notre application, les objets sont des verbes Français et les attributs principalement des cadres syntaxiques. Intuitivement, un concept est une paire $\langle O, A\rangle$ telle que les objets de O ont les attributs de A et inversement, les attributs de A sont valides pour tous les objets de O. En fait, nos concepts vont

regrouper précisément les ensembles de verbes qui partagent exactement le même ensemble de traits syntaxiques et/ou sémantiques.

Plus formellement, un contexte formel $\mathcal{K}$ est un triplet $\langle \mathcal{O}, \mathcal{A}, R \rangle$ tel que $\mathcal{O}$ est un ensemble d'objets, $\mathcal{A}$ un ensembles d'attributs et R la relation définit sur $\mathcal{O} \times \mathcal{A}$. Partant d'un tel contexte, un concept est une paire $\langle O, A \rangle$ telle que

$$O = \{o \in \mathcal{O} \mid \forall a \in A.\ (o, a) \in R\}$$
$$\text{et inversement}$$
$$A = \{a \in \mathcal{A} \mid \forall o \in O.\ (o, a) \in R\}.$$

Deux opérateurs, dénotés par $'$, relient les ensembles des sous-ensembles d'objets $2^{\mathcal{O}}$ et d'attributs $2^{\mathcal{A}}$ comme suit :

$$' : 2^{\mathcal{O}} \rightarrow 2^{\mathcal{A}},$$
$$X' = \{a \in \mathcal{A} \mid \forall o \in X.\ (o, a) \in R\}$$

L'opérateur $'$ est définit d'une façon duale pour les attributs. Pour un concept formel $\langle O, A \rangle \in O \times A$ nous avons $O' = A$ et $A' = O$. O est appelé l'*extension* et A l'*intension* du concept formel.

Un concept $C1 = \langle O1, A1 \rangle$ est *plus petit* qu'un autre concept $C2 = \langle O2, A2 \rangle$ (ce qui s'écrit $C1 \leq C2$) ssi $O1 \subseteq O2$ et $A1 \supseteq A2$.

L'ensemble des concepts formels d'un contexte $\mathcal{K}$ avec l'ordre $\leq$ forment un treillis complet appelé $\mathbb{K}$, le treillis de concepts de $\mathcal{K}$. Plus précisément, pour chaque sous-ensemble de concepts, il existe toujours un sous-concept maximal commun unique et un super-concept minimal commun unique.

Pour l'acquisition d'une classification verbale associant groupes de verbes à des ensembles de cadres syntaxiques nous appliquons FCA à un contexte formel, où les objets sont des verbes français et les attributs principalement les cadres syntaxiques associés à ce verbes par le lexique fusionné décrit en Section F.1.1.1. Ces expériences sont basées sur l'ensemble de données **vn_restricted**, qui est constitué approximativement de 2200 verbes français présents dans les 11 classes Verbnet de la référence V-gold. Nous construisons des classifications en utilisant les ensembles de traits présentés en Section F.1.1 : les cadres de sous-catégorisation seuls (**scf**) ou combinés avec des traits syntaxiques (**synt**) et/ou sémantiques (**sem**).

Nous commençons nos explorations en utilisant comme traits les cadres syntax-

iques seulement. A l'aide de l'outil Galicia[4] nous créons un treillis de concepts à partir du contexte formel $\langle V, F, R\rangle$ tel que :

- V est l'ensemble de verbes présents dans les classes traduites ou dans la référence V-gold et dans notre lexique de sous-catégorisation. Nous ne prenons pas en compte les verbes à un seule cadre de sous-catégorisation.
- F est l'ensemble de cadres syntaxiques présents dans le lexique de sous-catégorisation, et
- R la relation tel que $(v, f) \in R$ ssi le lexique de sous-catégorisation associe le verbe v au cadre f.

Ce contexte formel comprend 2091 objets (verbes) et 238 attributs (cadres) et le treillis résultant est constitué de 12 802 concepts formels. Clairement, tous ces concepts ne représentent pas de classes de verbes pertinentes. Le but des classes verbales étant la factorisation et généralisation de l'information sur les verbes, des classes avec très peu de verbes (1 ou 2) et peu de cadres sont moins intéressantes de ce point de vue, d'où la nécessité de filtrer le treillis de concepts pour ne garder que les concepts les plus pertinents par rapport à notre objectif. Pour ce filtrage nous nous basons sur des *indices pour la sélection des concepts* qui ont été proposés et analysés dans [Klimushkin *et al.*, 2010]. Dans la suite nous explorons lequel de ces indices est le plus performant dans le contexte de notre application. Nous cherchons à déterminer l'indice ou la combinaison d'indices nous permettant de sélectionner les groupes de verbes qui nous semblent les plus cohérents d'un point de vue syntaxique et sémantique. Pour cela, nous procédons de la manière suivante. Nous sélectionnons, pour chaque indice ou combinaison d'indices les 1500 concepts avec l'indice le plus élevé. De ces concepts nous ne gardons que ceux, qui, après alignement, correspondent le mieux aux classes Verbnet traduites. Ainsi, ces concepts sont associés avec les rôles thématiques des classes Verbnet et peuvent être comparés à une référence (par exemple V-gold). Nous considérons que l'indice de sélection ou la combinaison d'indices la plus adapté à notre application est celui (ou celle) qui donne les meilleures résultats par rapport à la référence.

Nous introduisons maintenant brièvement les indices de sélection de concepts avant de présenter les résultats de nos expériences.

[Klimushkin *et al.*, 2010] proposent trois indices pour sélectionner les concepts les plus pertinents dans un treillis basé sur des données bruitées : Les indices de *stabilité, séparation* et de *probabilité* d'un concept.

[4] `http://www.iro.umontreal.ca/~galicia/`

La stabilité d'un concept $C = (V, F)$ est la proportion des sous-ensembles de l'extension V ayant le même ensemble d'attributs F que V :

$$\sigma((V,F)) = \frac{|\{A \subseteq V \mid A' = F\}|}{2^{|V|}}. \tag{1}$$

Intuitivement, un concept plus stable est moins dépendant d'un objet particulier dans son extension et de ce fait est moins influencé par des données aberrantes ou autres données inexactes.

La séparation d'un concept est une mesure indiquant la couverture réalisée par les objets et attributs d'un concept : combien d'objets sont couverts par l'extension par rapport à la totalité des objets et combien d'attributs sont couvert par l'intension :

$$\mathfrak{s}((V,F)) = \frac{|V|\,|F|}{\sum_{v\in V}|\{v\}'| + \sum_{f\in F}|\{f\}'| - |V|\,|F|}. \tag{2}$$

Un indice élevé suggère que les verbes et simultanément les cadres regroupés par ce concept sont particulièrement pertinents, comparés aux autres groupements possibles de verbes et de cadres. Contrairement à la stabilité, qui se calcule soit pour les objets, soit pour les attributs, la séparation concerne simultanément les objets et attributs.

La probabilité d'un concept. Pour un attribut $a \in A$, l'ensemble des attributs, dénotons par p_a la probabilité d'un objet d'avoir l'attribut a. En pratique, ceci représente la proportion d'objets ayant a : $p_a = \frac{|\{a\}'|}{|O|}$, où O est l'ensemble d'objets. Pour un ensemble $B \subseteq A$, p_B est défini comme la probabilité d'un objet arbitraire d'avoir tous les attributs de B : $p_B = \prod_{a\in B} p_a$. Notons que cette formulation présuppose l'indépendance réciproque des attributs. À partir de cette formule, et en dénotant $n = |O|$, la probabilité de B d'être un ensemble fermé est exprimé comme suit :

$$p(B = B'') = \sum_{k=0}^{n} p(|B'| = k, B = B'') \tag{3}$$

$$= \sum_{k=0}^{n} \left[\binom{n}{k} p_B^k (1-p_B)^{n-k} \prod_{a\notin B} (1-p_a^k) \right] \tag{4}$$

Une valeur basse de $p(B = B'')$ suggère qu'il est peu probable que la combinaison d'attributs B est l'intension d'un concept uniquement par chance. Cependant, ce raisonnement est basé sur l'indépendance des attributs qui n'est pas donnée dans

	couv.	P	R	F2
stabilité	39.88	18.96	32.55	26.27
séparation	34.25	28.37	21.52	23.41
probabilité	35.53	26.60	20.73	22.38
sans filtrage	100	12.30	60.96	26.30

Tableau 3: Scores F2 et couverture pour les indices de stabilité, séparation et probabilité.

notre application.

Calculer les indices de sélection de concepts. Bien que la complexité du calcul de l'indice de *stabilité* soit de #P, il a été montré dans [Roth *et al.*, 2006] que si le treillis de concepts est connu, la stabilité peut être calculée efficacement par une approche transversale ascendante et c'est cet algorithme que nous avons utilisé dans nos calculs. L'indice de *séparation* peut être calculé facilement dans un temps de $\mathcal{O}(|O|+|A|)$, où O et A représentent l'ensemble d'objets et respectivement d'attributs. En ce qui concerne la *probabilité*, [Klimushkin *et al.*, 2010] montre que calculer la probabilité d'un seul concept nécessite un nombre très élevé d'opérations de multiplications : $\mathcal{O}(|O|^2 \cdot |A|)$. En raison de cette complexité nous avons eu recours à une approximation :

$$p(B = B'') = \sum_{k=0}^{40} \left[\binom{n}{k} p_B^k (1-p_B)^{n-k} \prod_{a \notin B} (1 - p_a^k) \right] \tag{5}$$

$$+ 1 - F(40; n, p_B). \tag{6}$$

où $F(k; n, p) = \sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-1}$ désigne la fonction de distribution cumulative de la distribution binomiale[5].

Nous procédons maintenant à l'évaluation de ces indices, dans le cadre expérimental précisé plus haut. Le Tableau 3 montre les résultats de ces premières expériences par comparaison à la référence, où nous évaluons la performance de chaque indice pris séparément. La précision et le rappel sont calculés par rapport aux associations ⟨verbe, grille thématique⟩ d'une part induit par la référence (`Ref`) et les classes

[5]Source Wikipedia: `http://en.wikipedia.org/wiki/Binomial_distribution`

sélectionnées (`Classif`) d'autre part.

$$R = \frac{|\{\langle \text{verbe, grille thématique} \rangle \in \text{Ref} \cap \text{Classif}\}|}{|\{\langle \text{verbe, grille thématique} \rangle \in \text{Ref}\}|}$$
$$P = \frac{|\{\langle \text{verbe, grille thématique} \rangle \in \text{Ref} \cap \text{Classif}\}|}{|\{\langle \text{verbe, grille thématique} \rangle \in \text{Classif} \mid \text{verbe } \in \text{Ref}\}}$$
$$F2 = \frac{3 * P * R}{2 * P + R}$$

Comme pour notre tâche le rappel est plus important que la précision, nous utilisons ici la mesure F2, qui donne plus de poids au rappel. La référence est composée des exemples positifs des associations ⟨verbe, grille thématique⟩ que nous avons utilisé pour l'apprentissage des classes traduites. Nous avons choisi cette référence plutôt que V-gold parce que ici nous évaluons les indices de sélections des concepts : Nous cherchons à déterminer l'index qui permet de sélectionner les concepts correspondants au mieux aux classes traduites. Les résultats montrés dans le Tableau 3 confirment les observations de [Klimushkin *et al.*, 2010] : La stabilité à elle seule permet de sélectionner des concepts aussi proche de la référence qu'une marge supérieure (les résultats obtenu sans filtrage). Les résultats des filtrages par séparation et probabilité sont beaucoup moins concluants. Comme nous ne sélectionnons qu'un petit nombre de concepts nous devons nous assurer que les classes sélectionnées couvrent une proportion satisfaisante des verbes à classifier. De ce point de vue les résultats pour les ensembles de classes sélectionnés sont décevants : le nombre de verbes éliminés par le filtrage est trop important (autour de 30% jusqu'à presque 40%).

Suivant les suggestions de [Klimushkin *et al.*, 2010], nous étudions dans la suite si un score issue d'une combinaison linéaire des trois indices de sélection permet un filtrage résultant en une meilleure couverture sans que la mesure F2 soit trop détériorée. Plus précisément, nous cherchons les coefficients k_i tels que les concepts avec le meilleure score :

$$k_1 \cdot \text{stability} + k_2 \cdot \text{separation} - k_3 \cdot \text{probability}$$

présentent une bonne mesure F2 et couvrent une proportion plus importante des verbes classifiés. En appliquant le même cadre expérimental que pour les indices seuls nous avons trouvé que la meilleure combinaison linéaire est la somme des indices de stabilité et séparation. La probabilité ne s'est pas montrée utile dans cette tâche, soit pour avoir présupposé à tort l'indépendance des attributs, soit parce que notre approximation est trop inexacte. Par conséquent, nous considérons que la meilleure stratégie pour filtrer le treillis est de sélectionner les 1500 concepts avec la somme des indices de stabilité et probabilité la plus élevée.

Attribuer des grilles thématiques aux concepts. Pour associer les concepts FCA avec des rôles thématiques nous les alignons avec les classes Verbnet traduites. Pour chaque classe traduite nous identifions parmi les 1500 classes sélectionnées celles avec la meilleure F-mesure entre les verbes partagés. Pour une classe Verbnet traduite C_{VN} et l'extension (ensemble de verbes) d'un concept FCA C_{FCA}, précision (P), rappel (R) et F-mesure (F) sont calculés comme suit :

$$R = \frac{|C_{\mathrm{VN}} \cap C_{\mathrm{FCA}}|}{|C_{\mathrm{VN}}|}, P = \frac{|C_{\mathrm{VN}} \cap C_{\mathrm{FCA}}|}{|C_{\mathrm{FCA}}|}, F = \frac{2RP}{R+P}$$

La classe traduite Verbnet est alors associée à ce concept FCA. De cette manière les verbes du concept FCA sont de fait associés aux rôles thématiques de la classe Verbnet. Comme nous l'avons montré précédemment, les classes traduites sont obtenues en utilisant différentes méthodes.

Le but de la série d'expériences suivante est double. Le premier objectif est de déterminer la méthode de traduction la plus adaptée à notre application. Le deuxième est d'explorer la performance des ensembles de traits utilisés pour la classification. Nous nous situons dans le même cadre d'expérimentation qu'auparavant : En utilisant FCA et différents ensembles de traits nous construisons des treillis qui sont ensuite filtrés par la somme des indices de stabilité et séparation. Les classifications résultantes sont alignées avec les classes traduites obtenues à l'aide de différentes méthodes et les classes sont ainsi enrichies de rôles thématiques. Les combinaisons de verbes et rôles thématiques issues de ces associations sont comparées à la référence V-gold en terme de précision, rappel et F-mesure. Pour résumer, les ensembles de traits évalués ici sont : **scf** (cadres de sous-catégorisation seulement), et cadres de sous-catégorisation combinés avec des traits syntaxiques (**synt**) et/ou sémantiques (**sem**). Les classes Verbnet traduites sont constituées en fonction de la fréquence des traductions (**median**) ou par apprentissage automatique **svm**. En fonction de l'attribution des verbes en classes à partir des probabilités résultant de l'apprentissage automatique nous obtenons les ensembles de classes traduites suivantes: **svm**, **svm-best** et **svm-median**. Ces expériences ont montré que l'ensemble de traits le plus performant (résultant en la meilleure F-mesure de 37.47) est l'ensemble **scf-sem**. Ce résultat a été obtenu en utilisant les classes traduites **svm**.

Finalement, nous avons montré dans [Falk *et al.*, 2010] et [Falk and Gardent, 2010] où nous nous sommes basés sur la même plate-forme, que les regroupements de verbes présentent une capacité importante de factorisation. De plus, une évaluation partielle qualitative suggère qu'une telle classification produit des classes cohérentes d'un point de vue syntactique et sémantique.

1.3.2 IGNGF

La deuxième technique de clustering que nous utilisons pour l'acquisition de classes verbales à la Verbnet s'appuie sur un algorithme incrémental de gaz neuronal croissant basé sur l'étiquetage des clusters par maximisation de vraisemblances. Les travaux présentés ici, ayant pour but d'appliquer cette méthode à l'acquisition de classes verbales et leur étiquetage avec des cadres de sous-catégorisation et des rôles thématiques ont été menés en collaboration avec Jean-Charles Lamirel. Nous commençons par introduire les grands principes de cette méthode.

IGNGF étant une méthode incrémentale, elle part d'une situation initiale où chaque point (à classifier) est considéré comme un cluster. À chaque itération, où tous les points sont parcourus, un point est connecté aux clusters les plus "proches". Simultanément, ce point influe sur le clustering actuel en renforçant les connections entre les clusters les plus proches et en diminuant les connections aux autres clusters, considérés moins pertinents. La notion de proximité est basée sur une distance qui est calculée à partir des vecteurs de traits associés aux data points.

Contrairement à ses précurseurs (NG - neural gas, GNG - growing neural gas et IGNG - incremental growing neural gas), la mesure de distance est remplacé dans IGNGF par la maximisation des vraisemblances des traits. La maximisation de ces vraisemblances est une métrique qui fait ressortir les clusters avec une F-mesure des traits maximale. La F-mesure des traits (FF - *Feature F-measure*) est la moyenne harmonique du rappel des traits (FR - *feature recall*) et de la précision des traits (FP - *feature precision*), définis comme suit :

$$FR_c(f) = \frac{\sum\limits_{v \in c} W_v^f}{\sum\limits_{c' \in C} \sum\limits_{v \in c'} W_v^f}, \qquad FP_c(f) = \frac{\sum\limits_{v \in c} W_v^f}{\sum\limits_{f' \in F_c, v \in c} W_v^{f'}}$$

où W_x^f représente le poids du trait f pour l'élément x et F_c désigne l'ensemble de traits associés aux verbes du cluster c. Un trait est alors considéré maximal pour un cluster c ssi sa F-mesure est plus élevée pour ce cluster que pour tout autre cluster. Finalement, la feature F-mesure FF_c d'un cluster est la moyenne des features F-mesures des traits maximaux pour c :

$$FF_c = \frac{\sum\limits_{f \in F_c} FF_c(f)}{|F_c|}$$

La méthode IGNGF s'est montrée particulièrement performante pour des tâches

de clustering sur des données relativement propre. Comme nos traits proviennent de ressources validées manuellement, elles peuvent être considérées plus propre que s'ils étaient extraits d'un corpus, et de ce fait se prêtent bien à une classification avec IGNGF.

L'objectif principal de cette méthode et de produire des groupes de verbes. Cependant, une particularité de IGNGF est d'être propice à l'étiquetage de ces clusters avec les traits considérés les plus caractéristiques pour le cluster par rapport (i) au clustering et (ii) à l'application. Ces techniques de maximisation des vraisemblances et de l'étiquetage des clusters se sont montrées efficaces pour la visualisation et la validation des clusters obtenus. Nous appliquons systématiquement ces procédures dans toutes nos expériences. L'étiquetage des clusters facilite l'interprétation des clusters en indiquant clairement l'association entre groupes de verbes et les traits les plus caractéristiques. D'autre part il favorise la création des classes verbales à la Verbnet en permettant d'associer des groupes de verbes à des cadres syntaxiques et des rôles thématiques.

Dans la partie suivante nous montrons comment IGNGF est employé pour l'acquisition de classes verbales Verbnet. Nous déterminons la configuration obtenant la meilleure classification, nous montrons un exemple d'un cluster et nous proposons une brève discussion qualitative. Toutes les classifications créées en utilisant IGNGF ont été calculées par Jean-Charles Lamirel.

Les objets à classifier sont les verbes Français de l'ensemble de données **vn_restricted** (2200 verbes Français dans 11 classes traduites). Les classes Verbnet sont traduites par la méthode **svm**. Nous effectuons la classification à partir des ensembles de données décrits en Section F.1.2 : ils sont constitués des cadres de sous-catégorisation extraits du lexique syntaxiques (**scf**) combinés (i) avec des traits syntaxiques (**synt**) et/ou sémantiques (**sem**) extraits des ressources lexicales Dicovalence et Ladl, comme pour FCA et (ii) avec les traits sémantiques dérivés des classes Verbnet traduites (**grid**) : un verbe Français a une classe Verbnet s'il est membre de la traduction de cette classe. Pour déterminer l'ensemble de traits le plus utile pour la constitution de classes verbales nous effectuons le clustering en nous appuyant sur chacun de ces ensembles de traits et comparons la classification résultante à la référence V-gold. Cette comparaison est effectuée sur la base des mesures de pureté (mPUR) et précision (ACC - accuracy) typiquement employées dans la recherche sur le clustering (notamment par [Sun *et al.*, 2010]). Le Tableau 4 montre les résultats de ces expériences. Les résultats sont meilleurs pour l'ensemble **grid-scf-sem** (F-mesure de 0.70). Comme dans le cas de FCA, les traits sémantiques (extraits des ressources lexicales et obtenus par traduction) se sont montrés utiles pour la constitution de

Ensemble de traits	Nbr. traits	Nbr. verbes	mPUR	ACC	F (Gold)	Nbr. classes
scf	220	2085	0.93	0.48	0.64	17
grid-scf	231	2085	0.94	0.54	0.68	13
grid-scf-sem	237	2183	0.86	0.59	**0.70**	13
grid-scf-synt	236	2150	0.87	0.50	0.63	14
grid-scf-synt-sem	242	2201	0.99	0.52	0.69	16
scf-sem	226	2183	0.83	0.55	0.66	23
scf-synt	225	2150	0.91	0.45	0.61	15
scf-synt-sem	231	2101	0.89	0.47	0.61	16

Tableau 4: Résultats (comparaison à V-gold) pour la méthode IGNGF, en variant les ensembles de traits. Ces ensembles sont constitués des cadres de sous-catégorisation combinés à des traits syntaxiques (*synt*) et sémantiques (*sem*) extraits des ressources lexicales Françaises et des traits dérivés des classes Verbnet traduites (*grid*).

C6- 14(14) [197(197)]
———-
Prevalent Label — = AgExp-Cause

0.341100 G-AgExp-Cause
0.274864 C-SUJ:Ssub,OBJ:NP
0.061313 C-SUJ:Ssub
0.042544 C-SUJ:NP,DEOBJ:Ssub
****
****
0.017787 C-SUJ:NP,DEOBJ:VPinf
0.008108 C-SUJ:VPinf,AOBJ:PP
...
[**déprimer 0.934345 4(0)] [affliger 0.879122 3(0)] [éblouir 0.879122 3(0)] [choquer 0.879122 3(0)] [décevoir 0.879122 3(0)] [décontenancer 0.879122 3(0)] [décontracter 0.879122 3(0)] [désillusionner 0.879122 3(0)] [**ennuyer 0.879122 3(0)] [fasciner 0.879122 3(0)] [**heurter 0.879122 3(0)] ...

Figure 2: Exemple d'un cluster produit avec la méthode IGNGF et l'ensemble de traits **grid-scf-sem**.

classes syntactico-sémantiques. L'exemple d'un cluster produit par cette procédure est montré dans la Figure 2. Notons que le cluster est visualisé en affichant les verbes et les traits maximaux par ordre décroissant de la feature F-mesure et les traits dont la F-mesure est en dessous d'une moyenne globale sont clairement démarqués.

Nous poursuivons ces investigations par une analyse préliminaire qualitative qui montre, que la technique IGNGF combinée aux approches d'étiquetage inhérentes, permet de produire une classification syntactico-sémantique cohérente. Nous avons effectué une analyse manuelle des clusters en examinant la pertinence sémantique du cluster (les verbes du cluster, partage-t-il une composante sémantique ?) et les associations des verbes, cadres syntaxiques et rôles thématiques induis par le clustering. Pour vérifier l'homogénéité sémantique nous identifions pour chaque cluster un ou plusieurs ensembles de rôles Verbnet caractérisant les verbes de ce cluster. Des

C0	speaking: babiller, bafouiller, balbutier SUJ:NP,OBJ:Ssub,AOBJ:PP *Jean bafouille à Marie qu'il l'aime*
C1	put: entasser, répandre, essaimer SUJ:NP,POBJ:PP,DUMMY:REFL Loc, Plural *Les déchets s'entassent dans la cour*
C2	hit: broyer, démolir, fouetter SUJ:NP,OBJ:NP Nhum *Ces pierres broient les graines* other_cos: agrandir, alléger, amincir SUJ:NP,DUMMY:REFL *les aéroports s'agrandissent sans arrêt*
C4	dedicate: s'engager à, s'obliger à, SUJ:NP,AOBJ:VPinf,DUMMY:REFL *Cette promesse t'engage à nous suivre*
C5	conjecture: penser, attester, agréer SUJ:NP,OBJ:Ssub *Le médecin atteste que l'employé n'est pas en état de travailler*
C6	amuse: déprimer, décontenancer, décevoir SUJ:Ssub,OBJ:NP SUJ:NP,DEOBJ:Ssub *Travailler déprime Marie* *Marie déprime de ce que Jean parte*

C7	other_cos: dégager, vider, drainer, sevrer judgement SUJ:NP,OBJ:NP,DEOBJ:PP *vider le récipient de son contenu* applaudir, bénir, blâmer, SUJ:NP,OBJ:NP,DEOBJ:Ssub *Jean blame Marie d'avoir couru*
C9	characterise: promouvoir, adouber, nommer SUJ:NP,OBJ:NP,ATB:XP *Jean nomme Marie présidente*
C10	amuse: agacer, amuser, enorgueillir SUJ:NP,DEOBJ:XP,DUMMY:REFL *Jean s'enorgueillit d'être roi*
C11	light: rayonner,clignoter,cliqueter SUJ:NP,POBJ:PP *Jean clignote des yeux* motion: aller, passer, fuir, glisser SUJ:NP,POBJ:PP *glisser sur le trottoir verglacé*
C12	transfer_msg: enseigner, permettre, interdire SUJ:NP,OBJ:NP,AOBJ:PP *Jean enseigne l'anglais à Marie*

Tableau 5: Correspondances entre clusters (produits par IGNGF), cadres syntaxiques et classes Verbnet.

13 clusters produits par le clustering, 11 ont pu être étiquetés. Ces clusters, ainsi que leurs étiquettes syntaxiques et sémantiques sont présentés dans le Tableau 5. Le tableau montre que quelques clusters regroupent plusieurs sous-classes et inversement, quelques classes Verbnet sont dispersées sur plusieurs clusters. Ceci n'est pas nécessairement injustifié cependant. D'une part rappelons que les verbes Verbnet ont été regroupés en fonction de leurs rôles thématiques, ce qui explique par exemple le regroupement dans le même cluster C2 des classes Verbnet other_cos-45.4 et hit-18.1 qui partagent les rôles thématiques Agent, Instrument, Patient. Les traits associés à ce cluster indiquent que les verbes membres sont transitifs, sélectionnent un objet concret et peuvent être pronominalisé, ce qui est correct pour la majorité des verbes dans ce cluster.

Nous avons également examiné si les étiquettes les plus importantes d'un cluster sont bien compatibles avec les verbes et les classe(s) sémantiques manuellement attribués aux clusters. Les associations montrées au Tableau 5 suggèrent que, d'une manière générale les cadres prévalents associés aux clusters correspondent bien à la

syntaxe des verbes membres.

Attribuer des cadres syntaxiques et des grilles thématiques aux clusters. Pour l'acquisition de classes verbales à la Verbnet il est nécessaire d'associer les clusters produits par IGNGF avec des ensembles de cadres de sous-catégorisation et des rôles thématiques. Comme nous l'avons vu, IGNGF associe chaque cluster avec les traits qui se sont montrés les plus importants à la constitution de la classification. Comme les cadres syntaxiques et les grilles thématiques sont parmi les traits utilisés dans le clustering, il est en principe possible de les employer également dans l'étiquetage des clusters, associant ainsi ces derniers avec des cadres syntaxiques et des rôles thématiques. En pratique, nous avons exploré la meilleure manière de générer ces associations en examinant la capacité des classifications produites d'appuyer une tâche d'annotation en rôles sémantiques. Cette évaluation sera présentée en détails en Section F.3, ici nous décrivons brièvement les résultats qui montre le processus d'étiquetage le plus efficace. Comme la référence V-gold que nous avons utilisée jusqu'à maintenant ne permet pas d'évaluer les associations ⟨verbe, cadre syntaxiques⟩, cette évaluation est basée sur la référence SRL gold, introduite en Section F.1.2. SRL gold est constituée de phrases d'un corpus français annoté en dépendances syntaxiques, que nous avons manuellement étiquetées en rôles thématiques Verbnet.

Pour l'association des clusters avec des **cadres syntaxiques**, nous avons expérimenté deux méthodes d'étiquetage. La première (**Fmax**) attribue aux clusters les cadres syntaxiques qui maximisent ce cluster (et qui, par conséquent, ont eu la contribution la plus importante à la constitution du cluster). La deuxième méthode (**Fpos**) associe à chaque cluster les cadres syntaxiques avec une Feature F-mesure au dessus d'un seuil définit globalement – la moyenne globale des Feature F-mesure des traits maximaux. Ces traits n'ont pas nécessairement contribué à la constitution des clusters mais sont potentiellement importants pour l'interprétation des groupes de verbes en tant que classes syntactico-sémantiques. La méthode **Fpos** s'est révélée plus efficace. En plus, une attention particulière a dû être accordée aux cadres syntaxiques très fréquents (par exemple le cadre transitive). En effet, comme ces cadres ne sont pas souvent discriminatifs, ils ne sont pas importants pour le clustering et leur Feature F-mesure est peu élevée.

En ce qui concerne l'attribution de **rôles thématiques** nous avons trouvé que les rôles thématiques maximaux induisent un rappel très bas. Nous avons donc opté pour une approche différente (la même que pour l'attribution de rôles thématiques aux concepts FCA). Chaque cluster est associé avec les rôles thématiques de la classe

Verbnet dont la traduction partage le plus grand nombre de verbes avec le cluster.

En résumé, l'application de la méthode IGNGF sera évaluée en Section F.3 dans les configurations suivantes. La classification est effectuée avec l'ensemble de traits **grid-scf-sem**, résultant en des regroupements de verbes. Ces groupes de verbes sont associés à des ensembles de cadres syntaxiques soit en choisissant les traits les plus importants pour la constitution de chaque cluster (**Fmax**), soit en sélectionnant les traits dont la Feature F-mesure dépasse un certain seuil global (**Fpos**). Des rôles thématiques sont attribués aux clusters soit en leurs associant des traits maximaux, soit en les alignant aux classes Verbnet traduites (par la méthode **svm**). Selon cette évaluation, la méthode d'étiquetage la plus performante est **Fpos** pour les cadres syntaxiques et l'alignement aux classes traduites pour les rôles thématiques.

2 Évaluation des classes sémantiques

Cette section aborde une évaluation sémantique des groupements de verbes produits à partir des ressources lexicales décrites en Section F.1.1 et en utilisant les méthodes présentées en Section F.1.3. Cette évaluation est basée sur une comparaison avec la référence V-gold. Comme cette référence ne permet qu'une évaluation des associations de verbes et rôles thématiques (et non pas des verbes et cadres syntaxiques), l'analyse présentée ici porte principalement sur la cohésion sémantique de ces classes verbales. Dans un premier temps nous utilisons la méthodologie d'évaluation proposée dans [Sun *et al.*, 2010], qui est souvent employée dans la recherche sur le clustering. Dans un deuxième temps, nous associons les groupements de verbes générés par les deux méthodes, FCA et IGNGF, avec des classes Verbnet traduites et leurs rôles thématiques. Cette étape peut introduire du bruit, mais elle permet d'associer les classes/clusters de verbes avec des grilles thématiques et de comparer ces associations à la référence.

2.1 Métriques d'évaluation

Pour évaluer les classes produites par les deux méthodes de classification nous employons deux métriques d'évaluation. La première est la pureté des clusters (mPUR, modified cluster purity), la précision (ACC, weighted class accuracy) et leurs F-mesure. La seconde est le rappel, la précision et la F-mesure des paires ⟨verbe, grille thématique⟩ dérivées de la classification par rapport à celles présentes dans la référence.

Pureté et précision des clusters. Pour calculer ces scores, on attribue à chaque cluster C une classe *prévalente* de la référence (prev(C)). prev(C) est la classe de la référence qui partage le plus grand nombre de verbes avec C. Un verbe est alors considéré correct si, dans la référence, il est associé à la classe prévalente du cluster dont il est membre. La pureté (mPUR) est définit comme suit[6] :

$$mPUR = \frac{\sum_{C \in \text{Clustering}, |\text{prev}(C)|>1} |\text{prev}(C) \cap C|}{\sum_{C \in \text{Gold}} \text{Verbes}_{\text{Clustering} \cap \text{C}}},$$

où $\text{Verbes}_{\text{C} \cap \text{Clustering}}$ est le nombre de verbes dans le cluster C de la référence et dans le clustering.

La précision (ACC) représente l'ensemble des verbes dans la référence et le clustering, associés à une classe de référence, comparés au nombre total de verbes dans la référence et le clustering. Pour calculer ACC, à chaque classe C_{Gold} de la référence est attribué un cluster dominant, le cluster dom(C_{Gold}) qui partage le plus grand nombre de verbe avec la classe de référence.

$$ACC = \frac{\sum_{C \in \text{Gold}} |\text{dom}(C) \cap C|}{\sum_{C \in \text{Gold}} \text{Verbes}_{\text{C}}}$$

Finalement, la F-mesure est la moyenne harmonique de mPUR et ACC.

Notons que les mesures mPUR et ACC comme utilisée dans [Sun *et al.*, 2010] ont du être adaptées aux classifications produites par FCA qui, contrairement à la référence et celles créées par IGNGF, sont recouvrantes. La différence est que pour une classification recouvrante, le dénominateur est le nombre total de verbes dans les classes et non pas simplement le nombre de verbes dans la référence, comme dans le cas d'une classification non-recouvrante.

Précision, rappel et F-mesure. Pour cette évaluation nous exploitons l'association implicite dans la référence de verbes à des grilles thématiques. Les classes de la référence sont identifiées à des classes Verbnet et ensuite à l'ensemble de rôles thématiques de celles dernières. De ce fait, chaque verbe de la référence est associé à un ou plusieurs ensembles de rôles thématiques et il est possible de comparer ces associations à celles dérivées de nos classifications. Pour cela les classes de nos classifications sont alignées aux classes traduites : un cluster C_{classif} est associé à la classe Verbnet (traduite) C_{VN} avec la meilleure F-mesure entre précision (P) et rappel (R),

[6] Clusters dont la classe prévalente n'a qu'un élément sont ignorés.

où rappel et précision sont définit comme suit :

$$R = \frac{|C_{\text{classif}} \cap C_{\text{VN}}|}{|C_{\text{VN}}|}, P = \frac{|C_{\text{classif}} \cap C_{\text{VN}}|}{|C_{\text{classif}}|}.$$

De cette façon chaque cluster est associé à une grille thématique et les paires ⟨verbe, grille thématiques⟩ induites peuvent être comparées à celles dérivées de la référence par les mesures classiques de précision, rappel et F-mesure. La précision est la proportion $(Classif \cap Gold)/Classif$ des paires ⟨verbe, classe Verbnet⟩ dérivées de nos classifications qui sont correctes. Le rappel est la proportion de paires ⟨verbe, classe Verbnet⟩ trouvées et la F-mesure et la moyenne harmonique de précision et rappel.

Cumulative Micro Precision (CMP). La micro précision cumulée est un nouvel index non-supervisé (introduit dans [Lamirel *et al.*, 2011a]) qui permet d'évaluer la qualité globale d'un résultat de clustering en utilisant les traits attribués à chaque cluster plutôt que d'avoir recours à une référence. Il a été montré dans [Lamirel *et al.*, 2008; Attik *et al.*, 2006] que ce type de métrique non-supervisée, basée sur l'étiquetage des clusters et la maximisation des vraisemblances peut être très utile pour identifier la stratégie de clustering la plus adaptée à l'application en question. Comme l'étiquetage des clusters et la maximisation des vraisemblances jouent un rôle important dans la construction de nos classes verbales, nous considérons ce score comme un indice de taille pour la cohérence de ces classes. En principe cette mesure est indépendante de la méthode de clustering, mais actuellement nous ne l'appliquons qu'à la méthode IGNGF. Le score CMP est définit comme suit :

$$CMP = \frac{\sum_{i=|C_{inf}|..|C_{sup}|} \frac{1}{|C_{i+}|^2} \sum_{c \in C_{i+}, f \in F_c} P_c^f}{\sum_{i=|C_{inf}|..|C_{sup}|} \frac{1}{|C_{i+}|}}$$

où C_{i+} représente le sous-ensemble des clusters du clustering C pour lesquels le nombre de données associées est supérieur à i, et $C_{inf} = argmin_{c_i \in C}|c_i|$, cad. la taille minimale des clusters, $C_{sup} = argmax_{c_i \in C}|c_i|$, cad. la taille maximale des clusters. La micro précision P_c^f d'un trait f et un cluster c est définit comme $P_c^f = \frac{|v_c^f|}{|V_c|}$, où v_c^f représente l'ensemble de verbes ayant le trait f et V_c l'ensemble de verbes dans c. Ce score permet de capturer la qualité du clustering d'une part par rapport à sa structure (ce qui est habituellement mesuré par macro précision et rappel) et

d'autre part par rapport aux traits maximaux (dont rendent compte habituellement les mesures de micro précision et rappel).

2.2 Cadre expérimental

Ces expériences sont effectuées à partir du jeu de données **vn_restricted** (cf. Section F.1.2). Pour *FCA*, le contexte formel est construit à partir des 2091 verbes et de l'ensemble de traits **scf-sem** (qui s'est montrés le plus efficace dans les expériences décrites en Section F.1.3.1). Le treillis résultant est filtré par la méthode décrite en Section F.1.3.1. Les concepts sélectionnés sont alignés aux classes Verbnet traduites **svm** et ne sont gardés que les concepts les plus proches des classes Verbnet. Ces concepts sont associés aux grilles thématiques des classes Verbnet alignées.

En ce qui concerne IGNGF, les objets à classifier sont les verbes de l'ensemble de donées **vn_restricted** et le clustering est basé sur l'ensemble de traits **grid-scf-sem**, qui s'est montré le plus performant dans les expériences décrites en Section F.1.3.2. Nous comparons la performance de cette méthode à un référentiel obtenu en appliquant la méthode de clustering K-means sur le même ensemble de verbes et de traits. Pour chaque méthode nous varions le nombre de clusters entre 1 et 30 pour déterminer la partition avec une F-mesure optimale est un nombre de classes proche de la référence (11 classes). Comme pour les concepts FCA, les clusters sont associés à des grilles thématiques en les alignant aux classes Verbnet traduites.

2.3 Résultats

Le Tableau 6 présente les résultats pour les deux classifications, par rapport aux mesures de pureté, précision et leurs F-mesure (Tableau 6a) et par rapport aux associations ⟨verbe, grille thématique⟩ produites (Tableau 6b).

En ce qui concerne d'abord l'évaluation basée sur pureté, précision et leurs F-mesure, une première observation est que la méthode IGNGF atteint une F-mesure supérieure à celle de [Sun *et al.*, 2010], qui ce situe autour de 65 pour des verbes à très haute fréquence (4000 instances). Les résultats ne sont pas vraiment comparable cependant, en raison des ressources utilisées : nos traits proviennent de ressources lexicales alors que ceux de [Sun *et al.*, 2010] sont extraits automatiquement d'un corpus. Malgré ces différences, Sun *et al.* ont également constaté un gain en performance à l'utilisation de traits sémantiques, une observation confirmée par nos expériences.

Pour les clusters IGNGF, les mesures non-supervisées indiquent une cohésion

(a) Pureté, précision et F-mesure pour les classifications créées avec FCA et IGNGF, comparé à la référence V-gold, proposée par [Sun *et al.*, 2010].

	Pureté (mPUR)	Précision (ACC)	F-mesure	CMP à l'opt. (13cl.)	Couverture
FCA	32.30	95.61	48.29		100
IGNGF	86.00	59.00	70.00	0.30	72
K-Means	88.00	57.00	70.00	0.10	88
[Sun *et al.*, 2010]			55-65.4		

(b) Précision, rappel et F-mesure pour les paires ⟨verbes, grille thématique⟩ dérivées des classifications FCA et IGNGF.

	Couverture (verbes)	Précision	Rappel	F
FCA	96.17	24.09	75.00	36.47
IGNGF	100.00	27.16	26.67	27.16

Tableau 6: Évaluation des classes verbales créées par FCA et IGNGF, par rapport aux mesures de pureté (mPUR), de précision (ACC) et de leurs F-mesure (a) et par une comparaison des associations ⟨verbe, grille thématique⟩ avec la référence V-gold (b). La colonne 4 dans (a) montre l'index de micro-précision cumulée et la dernière colonne donne le nombre de classes avec un trait maximal de type **grid**.

importante des clusters avec un nombre de clusters proche de la référence (13 vs. 10), une micro précision cumulée élevée (CMP = 30 vs. 10 pour K-means). Si K-means et IGNGF atteignent une F-mesure similaire, la micro-précision peu élevée pour K-means indique la disposition de cette méthode à produire des clusters de taille plus importante et plus hétérogènes.

Comparant IGNGF et FCA, le Tableau 6a montre que les performances de la méthode IGNGF en ce qui concerne ces mesures globales d'évaluation, sont beaucoup plus élevées que celles de FCA (20 points de différence en F-mesure). Aussi il est intéressant de constater que pour IGNGF la pureté est plus élevée que la précision alors que pour FCA ce rapport entre pureté et précision est inversé. Étant donné qu'une valeur élevée de la pureté indique une similitude structurelle, le clustering IGNGF se montre structurellement beaucoup plus proche du clustering référence, alors que la classification FCA présente des différences structurelles importantes. Ceci n'est pas trop surprenant cependant vu le caractère recouvrant de la classification FCA. D'autre part, les résultats en précision suggèrent qu'un nombre important de verbes sont groupés d'une manière similaire dans les classes FCA que dans la référence, ce qui semble moins vrai pour les clusters IGNGF.

Par contre, dans l'évaluation des paires ⟨verbe, grille thématique⟩ produites, la classification FCA s'est montrée plus performante dans la mesure où la F-mesure pour FCA dépasse celle obtenu pour IGNGF (de 10 points). Plus spécifiquement, alors que pour IGNGF précision et rappel sont proches, pour FCA la précision baisse de 3% mais le rappel dépasse celui d'IGNGF de 50%. Une raison pour cela est la nature recouvrante de la classification FCA selon laquelle un verbe peut être

associé à plusieurs grilles thématiques. La valeur basse de la précision suggère que beaucoup des associations ⟨verbe, grille thématique⟩ induites par les classes FCA sont incorrectes selon la référence. Cependant un certain nombre de ces associations peuvent être correctes mais simplement manquer dans la référence.

Pour conclure, l'évaluation présentée ici montre que les deux classifications obtenues par les méthodes FCA et IGNGF permettent de créer des groupes de verbes cohérents d'un point de vue sémantique. Cette cohérence est plus évidente pour la méthode IGNGF. Le clustering produit par cette méthode s'est montré plus performant que celui de [Sun *et al.*, 2010] en terme de pureté, précision et leurs F-mesure. La cohérence des classes FCA est montrée par le rappel élevée (75%) des paires ⟨verbe, grille thématique⟩ induites par la classification et comparée à celles dérivées de la référence. La F-mesure très basse résulte d'un niveau bas de précision mais qui n'est pas nécessairement du à des associations incorrectes.

3 Évaluation des classes syntactico-sémantiques

Cette section est dédiée à une évaluation des classifications obtenues avec FCA et IGNGF par rapport non seulement aux groupes de verbes créés mais aussi aux cadres syntaxiques et grilles thématiques associés. Nous évaluons la capacité de nos classifications d'associer les verbes avec des ensembles de cadres syntaxiques et rôles thématiques appropriés en comparant ces associations avec celles présentes dans la référence SRL gold, présentée en Section F.1.2. Cette évaluation est effectuée de deux façons. Premièrement, d'un point de vue plus global, les paires ⟨verbe, cadre syntaxique⟩ et ⟨verbe, grille thématique⟩ de la référence sont comparées à celles induites par les classifications. Deuxièmement, nous conduisons une évaluation axée sur une tâche, en utilisant les classifications dans une tâche d'étiquetage en rôles thématiques. Ceci permet une évaluation plus fine au niveau de l'interface syntaxe-sémantique dans la mesure où nous analysons la capacité de nos classifications de faciliter l'attribution de rôles thématiques à des verbes et leurs arguments syntaxiques.

3.1 Cadre expérimental

Dans ces expériences nous employons les méthodes FCA et IGNGF sur l'ensemble de données **vn_all**, présenté en Section F.1.2. Les verbes classifiés sont les 4260 verbes présents dans notre lexique de sous-catégorisation fusionné et l'ensemble de traits utilisé est **scf-sem** pour FCA et **grid-scf-sem** pour IGNGF, cad. les cadres de sous-catégorisation et les traits sémantiques extraits des ressources lexicales. De

plus, pour IGNGF, nous utilisons aussi les traits sémantiques basés sur les classes Verbnet traduites. Comme pour les expériences antérieures, la traduction des classes Verbnet a été effectuée avec la méthode **svm**.

FCA. En partant de ces données nous construisons le contexte formel comprenant 4260 verbes et 303 attributs (cadres de sous-catégorisations et traits sémantiques). Le treillis résultant est formé de 35 274 concepts qui sont filtrés comme décrit en Section F.1.3.1. Ces concepts, regroupant des ensembles de verbes et de cadres de sous-catégorisation, sont associés à des grilles thématiques en les alignant avec les classes Verbnet traduites en fonction des verbes membres communs. Seuls les concepts associés à une grille thématique sont gardé dans la classification finale. Le résultat de cette procédure est une classification regroupant 3994 verbes en 52 classes associées avec 32 cadres syntaxiques distincts et 61 ensembles de rôles thématiques. La Figure 3 montre un extrait de la classification obtenue, qui sera évaluée dans les sections suivantes. Par exemple, la classe 9109 contient 59 verbes (*eg. abaisser, accompagner, acheminer, apporter, avancer, baisser, balancer, bouger, cahoter, camionner, catapulter, charrier, colporter, coltiner, ...*) pouvant être utilisés dans la construction transitive simple (SUJ:NP,OBJ:NP) ou avec deux objets prépositionnels (SUJ:NP,OBJ:NP,POPJ:PP,POBJ:PP). D'un point de vue sémantique cette classe est associée à la grille thématique Agent-End-Start-Theme. Cette classe représente donc correctement des verbes de mouvement où un Agent déplace un Thème d'un endroit à un autre.

IGNGF Nous appliquons la méthode IGNGF sur les verbes du jeux de données **vn_all**, en utilisant l'ensemble de traits **grid-scf-sem** et obtenons ainsi des regroupements de verbes. Afin de rendre ce clustering similaire à Verbnet les clusters de verbes doivent être associés à des cadres de sous-catégorisation et des grilles thématiques. Pour produire ces associations nous appliquons les techniques d'étiquetage décrites en Section F.1.3.2. Comme les cadres syntaxiques et les grilles thématiques sont parmi les traits utilisés pour le clustering, il est possible de générer les associations recherchées sur la base des traits maximaux des clusters, mais, comme nous allons voir plus loin, cette approche ne donne pas des résultats satisfaisants. Dans les expériences décrites par la suite nous avons déterminé la technique d'étiquetage la plus adaptée à notre application, cad. celle où la classification résultante s'est montrée la plus utile dans une tâche d'attribution de rôles sémantiques. C'est cette classification qui est ensuite évaluée.

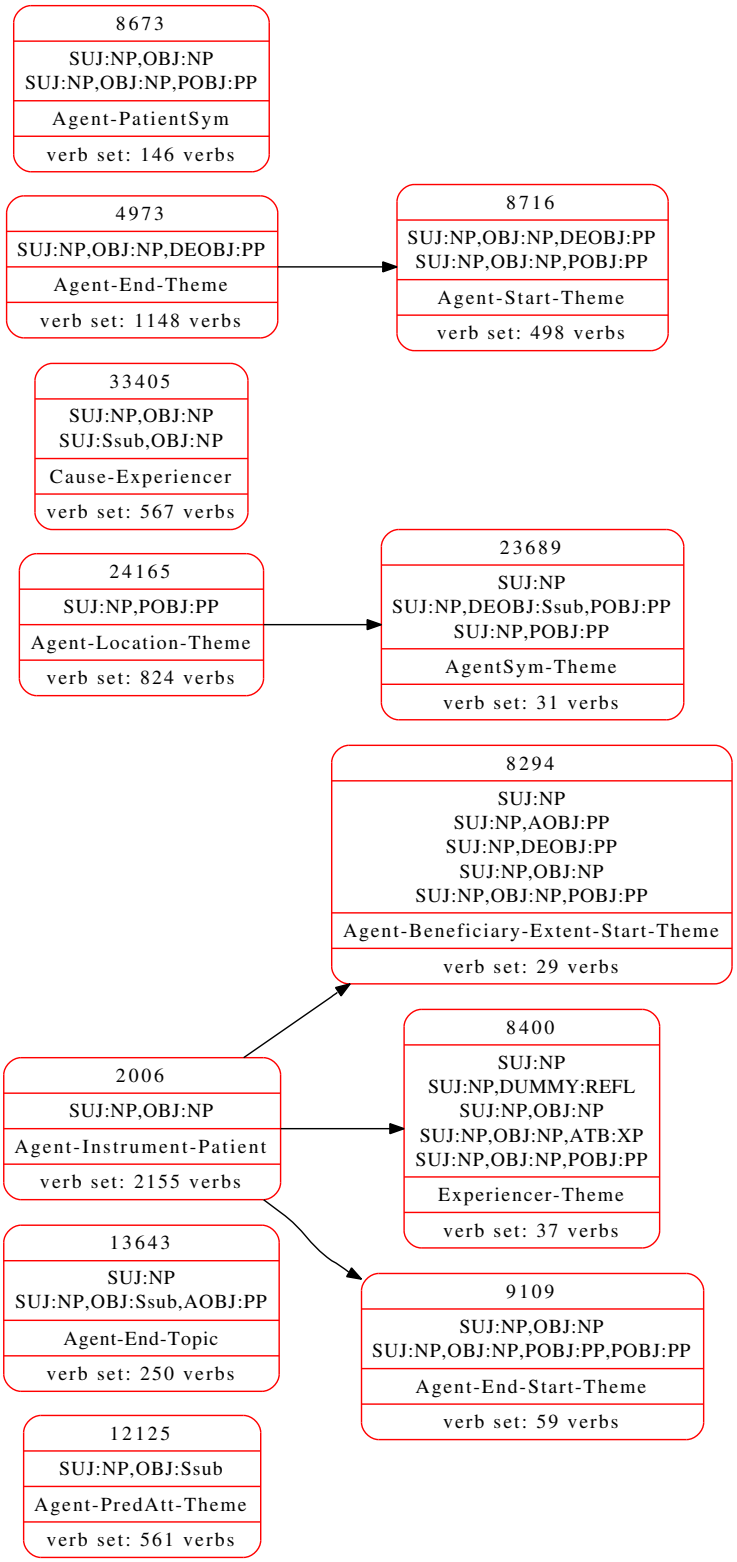


Figure 3: Extraits de la classification Verbnet associant verbes Français, cadres syntaxiques et grilles thématiques, obtenue par la méthode de classification FCA.

3.2 Scénarios d'évaluation

Cette section présente les deux scénarios d'évaluation que nous suivons pour évaluer les associations ⟨verbe, cadre syntaxique⟩ et ⟨verbe, grille thématique⟩ induites par nos classifications. Le premier scénario permet une évaluation sur un niveau plus global : nous vérifions dans quelle mesure les associations dérivées de nos classifications correspondent à celle de la référence SRL gold. Le deuxième schéma d'évaluation concerne l'interface syntaxe/sémantique. Nous étudions la contribution que nos classifications peuvent apporter à l'annotation automatique en rôles sémantiques des arguments syntaxiques des verbes du corpus SRL gold.

3.2.1 Niveau global.

Pour l'évaluation des associations des verbes avec des cadres syntaxiques et grilles thématiques nous comparons simplement les cadres syntaxiques et grilles thématiques présents dans la référence SRL gold à celles générées par la classification.

3.2.2 Interface Syntaxe/Sémantique.

Pour implémenter ce scénario d'évaluation, nous utilisons chaque classification pour étiqueter les paires ⟨verbe, argument syntaxique⟩ de la référence SRL gold avec des rôles thématiques Verbnet et comparons ensuite ces associations à celles produites manuellement. Pour associer les paires ⟨verbe, argument syntaxique⟩ à des rôles thématiques nous appliquons une méthode proposée par [Swier and Stevenson, 2005], qui emploient Verbnet pour créer un étiqueteur en rôles thématiques non-supervisé pour l'Anglais.

L'étiquetage en rôles thématiques présenté par [Swier and Stevenson, 2005] est effectué sur les phrases de FrameNet II qui à leur tour proviennent du British National Corpus [Clear, 1993]. Ces phrases sont analysées syntaxiquement, procédure qui résulte dans des associations de verbes à des cadres syntaxique de la forme SUBJ V OBJ. En s'appuyant sur Verbnet et en appliquant une procédure appelée **frame matching** sur ces ensembles ⟨verbe, cadre⟩, Swier and Stevenson identifient les associations ⟨verbe, argument syntaxique, rôle thématique⟩ non-ambiguë les plus probable. Le *frame matching* est appliqué à une paire ⟨verbe, cadre⟩ comme suit : le verbe de la paire est un membre d'une ou plusieurs classes Verbnet. Ces classes Verbnet donnent les cadres syntaxiques des verbes membres d'une part et les grilles thématiques d'autre part, qui reflètent la réalisation syntaxique des rôles thématiques dans ces cadres. Par exemple, la Figure 1 montre les cadres syntaxiques et grilles thématiques associés par Verbnet aux verbes de la classe *amuse-31.1*. On

grilles thématiques pour V	construction syntaxique		$\%\theta$	%SCF	Score
	SUBJ	OBJ			
Agent V	Agent		100	50	150
Agent V Theme	**Agent**	**Theme**	100	100	**200**
Instrument V Theme	**Instrument**	**Theme**	100	100	**200**
Agent V Recipient Theme	Agent	Theme	67	100	167

Tableau 7: Illustration de *frame matching*. Les arguments syntaxiques à étiqueter sont SUBJ et OBJ (pour un verbe V). D'après Verbnet, le verbe V peut être utilisé dans les grilles thématiques listées en première colonne. Les grilles thématiques sont marquées par la somme de la proportion d'arguments syntaxiques alignés à un rôle thématique (%SCF) et des rôles thématiques alignés à un argument syntaxique ($\%\theta$).

voit que la grille thématique **Cause, Experiencer** est une réalisation syntaxique du cadre transitif simple (NP V NP ou SUJ,OBJ). Ce sont ces grilles thématiques qui sont utilisées dans le processus de frame matching. Le cadre de la paire est aligné avec chacune de ces grilles thématiques et à chacun de ces alignements est attribué un score mesurant la correspondance entre cadre syntaxique et grille thématique. Le Tableau 5.5 illustre cette procédure. Des grilles thématiques présentées dans ce tableau, ne sont retenues que celles avec le score le plus élevé. De celles-ci nous dérivons les associations ⟨argument syntaxique, rôle thématique⟩ suivantes : SUBJ est associé aux rôles **Agent** et **Instrument** et OBJ à **Theme**. Comme seule la dernière association est non-ambiguë, ce n'est que celle-ci qui est retenue. Par conséquent, les instances ⟨V, OBJ⟩ du corpus sont étiquetées avec le rôle thématique **Theme**.

Dans l'étape suivante, les associations ⟨verbe, argument syntaxique, rôle thématique⟩ non-ambiguës sont utilisées pour calculer les probabilités P(rôle sémantique|argument syntaxique) qui ensuite interviennent dans la désambiguïsation des associations restées ambiguë lors du frame matching.

Pour évaluer leur méthode, Swier and Stevenson comparent les triplets ⟨verbe, argument syntaxique, rôle sémantique⟩ obtenus à ceux d'une référence de 1000 phrases où aux arguments syntaxiques issus de l'analyse syntaxique automatique ont été attribué manuellement des rôles thématiques. Ils parviennent à une F-mesure de 0.76 en n'appliquant que le frame matcher et de 0.83 en utilisant les probabilités résultant de l'annotation initiale générée par le frame matching.

Nous décrivons maintenant comment la méthode de [Swier and Stevenson, 2005] est adaptée à nos ressources Françaises. Nous avons vu que les ressources Anglaises utilisées par la méthode de Swier and Stevenson sont la classification Verbnet et des instances ⟨verbe, argument syntaxique⟩ extraites automatiquement d'un corpus. Notre approche est basée sur les classes verbales issues de nos classifications au lieu

6583: 50 verbes, 5 cadres, 1 ensemble de rôles thématiques	
cadres	SUJ:NP SUJ:NP,OBJ:NP SUJ:NP,OBJ:NP,AOBJ:PP SUJ:NP,OBJ:NP,DEOBJ:PP SUJ:NP,OBJ:NP,DUMMY:REFL
rôles thématiques	Agent-Beneficiary-Start-Theme (classe VN *steal-10.5*)
verbes	acheter, assurer, attendre, attirer, avancer, câbler, connaître, cracher, croire, découvrir, demander, défendre, dire, délivrer, donner, envier, faire, faucher, fermer, fixer, flanquer, fournir, gagner, indiquer, inspirer, jeter, masquer, passer, payer, piquer, porter, prendre, présenter, ramener, rappeler, rapporter, reconnaître, rejeter, rendre, reprendre, représenter, servir, sortir, souffler, tendre, tenir, tirer, trouver, vendre, voler

Tableau 8: Classe Verbnet française, obtenue par une classification FCA.

de Verbnet et les instances ⟨verbe, argument syntaxiques⟩ extraites de la référence française SRL gold à la place des instances Anglaises. L'étiquetage de ces instances françaises en rôles thématiques Verbnet peut ensuite être comparé aux annotations manuelles de la référence, une comparaison qui permet de mesurer l'apport de nos classes verbales à cette tâche simplifiée d'étiquetage en rôles thématiques.

Nos deux méthodes de classification produisent des classes verbales associées à des cadres syntaxiques et rôles thématiques, telles que montrées aux Tableau 8. Pour appliquer le frame matching de Swier and Stevenson il est nécessaire d'associer ces classes à des grilles thématiques. Ces grilles thématiques sont extraits des classes Verbnet avec les mêmes rôles thématiques. Par exemple, pour la classe 6583 représentée dans le tableau ce sont les grilles données par la classe Verbnet *steal-10.5* : Agent Theme, Agent Theme Start, Agent Theme Beneficiary, Agent Theme Start Beneficiary[7]. Comme dans [Swier and Stevenson, 2005], ces grilles thématiques sont alignées aux cadres syntaxiques de la classe verbale mais, contrairement au frame matching anglais les cadres syntaxiques n'ont pas été extraits d'un corpus mais ont été associés aux verbes par la classification. Comme pour l'anglais nous constituons des associations ⟨verbe, argument syntaxique, rôle thématique⟩ pour chaque verbe d'une classe et chacun des arguments syntaxiques associé à ce verbe par sa classe verbale. Pour une classification recouvrante comme FCA cette méthode est très restrictive. Nous avons également exploré une méthode moins restrictive où les associations ⟨verbe, argument syntaxique, rôle thématique⟩ sont créées sur la base des verbes et cadres syntaxiques d'une seule classe verbale à la fois. Ces associations sont ensuite utilisées de la même façon que dans [Swier and Stevenson, 2005] pour calculer des estimations de probabilité et propager l'annotation de base obtenue par

[7]Nous faisons abstraction des prépositions et de la position du verbe parce que ceux-ci ne correspondent très probablement pas aux homologues français

frame matching au reste du corpus.

Pour notre évaluation nous utilisons chacune de nos classifications, l'une obtenue par la méthode FCA l'autre par IGNGF pour produire deux systèmes d'étiquetage en rôles sémantiques (SRL). Le premier est basé sur le frame matching seulement et le deuxième sur le frame matching combiné aux estimations de probabilité. Ces systèmes sont appliqués aux instances ⟨verbe, argument syntaxique⟩ de notre corpus (dont la référence SRL gold fait partie), pour leur associer des rôles thématiques Verbnet. Notre évaluation consiste alors dans une comparaison de ces étiquettes aux annotations de la référence SRL gold. Elle montre dans quelle mesure les associations ⟨verbe, argument syntaxique, rôle thématique⟩ engendrées par nos classification correspondent à celles de la référence. Le deuxième type de système SRL, combinant le frame matching et certaines estimations probabilistes, permet d'estimer comment les associations issues uniquement à partir de données lexicales peuvent être complétées à l'aide de données provenant du corpus.

Les étiquettes produites sont comparées d'une part à la référence, d'autre part à des associations par défaut, selon le tableau suivant :

SUJ:NP	Agent
OBJ:NP, OBJ:Ssub, OBJ:VPinf	Theme
AOBJ:PP, AOBJ:VPinf	End
DEOBJ:PP, DEOBJ:VPinf	Start
POBJ:PP, POBJ:Ssub, POBJ:VPinf	Location
ATB:XP	PredAtt

3.3 Évaluation FCA

Cette section présente les résultats de l'évaluation de la classification obtenue avec la méthode FCA, conduite suivant les méthodologies décrites plus haut.

3.3.1 Niveau global.

Le Tableau 9 résume la comparaison plus globale au niveau des associations des verbes à des cadres syntaxiques et grilles thématiques. Nous commentons d'abord les associations ⟨verbe, cadre⟩, montrées au Tableau 9a. Le rappel est de 76.90%. Ce score est biaisé par les 42 (13.92%) associations ⟨verbe, cadre⟩ présents dans la référence mais manquant dans le lexique, qui par conséquent n'ont pas pu être générées par la classification. Le rappel ajusté est de 88.68%. Des 274 paires ⟨verbe, cadre⟩ dans la référence, 31 (11.31%) n'ont pas été générées par la classification. Ceci indique que les cadres associés aux classes verbales sont souvent peu informatifs. Cette supposition a été confirmée par l'observation que souvent les classes sont

(a) Associations entre verbes et **cadres syntaxiques**.

SRL gold	classif	SRL gold et lex	SRL gold et classif	SRL gold, lex, pas classif	SRL gold pas lex	Rappel	Rappel sans manquants en lex
316	16542	274	243	31	42	76.90	88.69

(b) Associations entre verbes et **grilles thématiques**.

SRL gold	classif	SRL gold et classif	Rappel
318	33525	280	88.05

Tableau 9: Évaluation sur un niveau plus global des associations entre verbes et cadres syntaxiques (a) et verbes et grilles thématiques (b) engendrées par la classification FCA (*classif*). Les associations aux cadres syntaxiques sont comparées aux annotations de la référence (*SRL gold*) et aux lexique (*lex*). Les associations aux grilles thématiques sont comparées aux associations de la référence seulement. Sont représentés des types, ie. le nombre de paires ⟨verbe, cadre⟩ ou ⟨verbe, grille thématique⟩ distinctes.

associées à peu de cadres très fréquents. Un moyen d'améliorer ces classifications serait donc de se concentrer plus sur les cadres, par exemple en filtrant les concepts par l'indice de stabilité des attributs (et non pas des objets comme cela est fait actuellement).

Puisque la référence ne contient qu'un nombre limité de phrases, la classification comprend beaucoup plus d'ensembles ⟨verbe, cadre⟩ que la référence, nous ne calculons donc pas la précision.

En ce qui concerne maintenant les grilles thématiques, des 318 paires ⟨verbe, grille thématique⟩ présentes dans la référence, 280 (88.05%) sont compatibles avec une paire ⟨verbe, ensemble de rôles thématiques⟩ engendrée par la référence, dans la mesure où les rôles de la référence représente un sous-ensemble d'un ensemble de rôles associés au même verbe par la classification. 38 paires (presque 12%) de la référence n'ont pas été produites par la classification. Nous avons identifié les raisons suivantes, ayant pu occasionnées ce décalage : erreurs d'annotation, confusions entre rôles thématiques (par exemple le rôle Location est souvent confondu avec les rôles Source ou Destination qui peuvent aussi désigner un rôle locatif), pas de classe Verbnet correspondante ou pas de rôle approprié et finalement dans 5 cas, le cadre syntaxique dans la référence manquait dans la classification. Ces associations manquantes pourraient aussi s'expliquer par l'approche d'alignement utilisée pour associer les classes FCA avec des rôles thématiques. Cette approche, basée sur une moyenne harmonique entre verbes partagés et taille des classes, résulte souvent dans l'alignement de classes FCA hétérogènes et de grande taille à des classes Verbnet très petites et précises. Par conséquent, une méthode plus élaborée pour l'attribution de rôles thématiques serait très souhaitable.

3.3.2 Interface syntaxe/sémantique.

Nous présentons ici les performances de l'étiqueteur en rôles thématiques Verbnet, construit à partir de la classification FCA suivant la méthodologie de [Swier and Stevenson, 2005], adaptée à nos ressources. Nous avons étudié l'impact de trois facteurs sur cet étiquetage :

Frame matching. Nous appliquons une méthode de frame matching plus ou moins restrictive (cf. Section F.3.2.2).

Corpus. Le frame matching est appliqué aux paires ⟨verbe, argument syntaxique⟩ du SRL gold seulement, ou à toutes les données dont nous disposons (le corpus P7 en entier).

Représentation des arguments syntaxiques. Les arguments syntaxiques sont représentés d'une manière plus ou moins détaillée.

Les étiquettes produites sont comparées en terme de F-mesure entre précision et rappel, d'une part à celles de la référence SRL gold, et d'autre part à un étiquetage par défaut.

Le Tableau 10 résume les résultats pour l'étiquetage obtenu dans ces différentes configurations. Nous constatons que la F-mesure résultante est dans toutes les configurations beaucoup plus basse que celle basée sur les associations par défaut (d'approximativement 20 points). Cette valeur est du principalement au rappel : un nombre important (autour de 63% de paires ⟨verbe, argument syntaxiques⟩) n'a pas pu être étiquetées par le frame matching. Ceci s'explique principalement par le fait qu'un nombre important de paires ⟨verbe, cadre⟩ n'a pas été générées par la classification et confirme ce qui a été constaté lors de l'évaluation au niveau global. Cependant, des 42% des associations qui ont pu être étiquetées, autour de 68% ont été attribuées au bon rôle thématique, ce qui représente une précision supérieure de 3% à l'étiquetage par défaut.

Les paires ⟨verbe, argument syntaxique⟩ manquantes dans la classification expliquent le gain relativement petit en performance lors de l'extension du modèle par estimations probabilistes. Comme certains arguments syntaxiques ne sont pas générés par la classification, ils ne sont jamais étiquetés et ne permettent pas une estimation probabiliste. Malgré cela, les estimations probabilistes permettent d'étiqueter la moitié des paires ⟨verbe, argument syntaxique⟩ qui n'ont pas pu été étiquetées par le frame matching, sans baisse importante de la F-mesure.

(a) Résultats en utilisant une représentation détaillée des arguments syntaxiques (fonctions et plusieurs catégories phrastiques).

		frame matching restrictif			frame matching moins restrictif		
	BL	FAm	FAm+prob	FAm+p7+prob	FAm1	FAm1+prob	FAm1+p7+prob
F	65.21	40.02	40.03	38.89	37.94	58.48	58.35
Préc.	65.21	68.00	48.29	45.64	70.24	59.77	59.52
non étiqu.	0	58.31	29.38	22.69	63.00	4.30	3.86

(b) Résultats en utilisant une représentation moins détaillée des arguments syntaxiques (fonctions, catégories nominales et phrastiques uniquement).

		frame matching restrictif			frame matching moins restrictif		
	BL	FAm	FAm+prob	FAm+p7+prob	FAm1	FAm1+prob	FAm1+p7+prob
F	65.21	42.92	54.41	42.22	37.85	56.46	56.46
Préc.	65.21	70.40	63.99	47.84	69.61	59.04	59.04
non étiqu.	0	56.14	26.05	21.03	62.66	8.38	8.38

(c) Résultats en utilisant une représentation simple des arguments syntaxiques (fonctions seulement).

		frame matching restrictif			frame matching moins restrictif		
	BL	FAm	FAm+prob	FAm+p7+prob	FAm1	FAm1+prob	FAm1+p7+prob
F	65.21	41.35	42.02	38.89	37.79	56.87	57.50
Préc.	65.21	77.84	69.87	44.60	69.81	61.30	63.01
non étiqu.	0	63.83	57.00	22.69	62.89	13.48	16.09

(d) Résultats (approximatifs) de [Swier and Stevenson, 2005].

	BL	FAm	FAm+prob
F	74	76	83
non étiqu.	0	38% (16 amb.)	

Tableau 10: Résultats de l'étiquetage obtenu : en utilisant une méthode plus (*FAm*) ou moins (*FAm1*) restrictive de frame matching; en étiquetant les paires ⟨verbe, argument syntaxique⟩ de SRL gold seulement (*FAm+prob*, *FAm1+prob*) ou toutes les paires du corpus P7 (*FAm+p7+prob*, *FAm1+p7+prob*). La représentation des arguments syntaxiques et plus (a) ou moins ((b) et (c)) détaillée. Les associations ⟨verbe, argument syntaxique, rôle thématique⟩ sont comparées à celles du SRL gold. Le Tableau 10d montre à titre indicatif les résultats de [Swier and Stevenson, 2005].

Appliquer la méthode de frame matching moins restrictive fait légèrement baisser la F-mesure de l'étiquetage basé sur le frame matching uniquement, mais entraîne une meilleure performance du modèle étendu par estimation des probabilités (en termes de F-mesure et de paires ⟨verbe, argument syntaxique⟩ étiquetées). Ce qui indique la cohérence des associations entre classes verbales et cadres syntaxiques.

En utilisant un ensemble plus important de données (toutes les instances ⟨verbe, argument syntaxique⟩ du corpus P7 et pas uniquement celles de la référence), la F-mesure baisse légèrement (approx. 3 points) mais beaucoup plus d'instances ⟨verbe, argument syntaxique⟩ peuvent être étiquetées. Cet effet est le plus prononcé pour la représentation la plus détaillée d'arguments syntaxiques, la représentation qui a été

utilisé pour l'annotation du corpus P7.

Ce qui nous avait amené a explorer l'impact de la représentation syntaxique est le fait que les expériences de Swier and Stevenson sont basées sur une représentation d'arguments syntaxiques plus simple (il s'agit des fonctions sujet, objet, objet indirect et objet prépositionnel). Nos expériences montrent que les arguments syntaxiques représentés d'une façon plus détaillée (fonctions et catégories) permettent un étiquetage en rôles thématiques plus précis. Ils permettent de compenser partiellement la perte d'information due à nos ressources construites automatiquement et par conséquet moins précises.

3.4 Évaluation IGNGF

Comme nous l'avons décrit en Section F.1.3.2, la méthode IGNGF produit des clusters (non-recouvrants) des verbes français présents dans nos lexiques en s'appuyant sur les traits syntaxiques et sémantiques extraits des mêmes ressources lexicales. Ces traits consistent principalement des cadres de sous-catégorisation, mais également de traits sémantiques dérivés des classes Verbnet traduites (ensemble de traits **grid-scf-sem**). Pour associer ces clusters à des ensembles de cadres de sous-catégorisation et à des rôles thématiques Verbnet (issus des classes Verbnet), nous employons les techniques d'étiquetage présentées en Section F.1.3.2, Plus précisément, il s'agit des approches suivantes :

Fmax Chaque cluster est associé aux cadres de sous-catégorisation qui maximise ce cluster.

Fpos Nous associons à chaque cluster les traits (cadres de sous-catégorisation) dont la F-mesure dépasse un seuil global.

theta-1 Nous associons chaque cluster aux rôles thématiques présents dans les étiquettes attribuées à ce cluster.

theta-2 Nous associons chaque cluster aux rôles thématiques des classes Verbnet traduites maximisant la F-mesure des verbes membres partagés [8].

Pour trouver la manière la plus efficace d'associer les clusters de verbes avec des cadres de sous-catégorisation et grilles thématiques, nous créons des classifications

[8] Si C_{VN} est une classe Verbnet traduite et C_{IGNGF} l'ensemble de verbes dans un cluster IGNGF, le rappel = $\frac{|C_{\text{VN}} \cap C_{\text{IGNGF}}|}{|C_{\text{VN}}|}$, la précision = $\frac{|C_{\text{VN}} \cap C_{\text{IGNGF}}|}{|C_{\text{IGNGF}}|}$ et la F-mesure est la moyenne harmonique entre précision et rappel : $\frac{2 * \text{precision} * \text{rappel}}{\text{precision} + \text{rappel}}$.

	%total (Rappel)	%labeled (Préc.)	F	%not labeled
theta-1	13.01	79.36	22.35	83.61
theta-2	30.26	76.72	43.41	60.55
Fmax	30.26	76.72	43.41	60.55
Fpos	47.02	72.53	57.05	35.17

Tableau 11: Résultats SRL obtenus en associant les clusters de verbes aux rôles thématiques (i) par les traits issues du clustering (ligne *theta-1*) et (ii) alignement avec les classes Verbnet traduites (ligne *theta-2*). Les associations aux cadres syntaxiques sont obtenues (i) en utilisant les traits maximaux des clusters (ligne *Fmax*) et (ii) en sélectionnant les traits dont la F-mesure est au dessus d'un seuil global (ligne *Fpos*). Le système d'étiquetage en rôle thématique est basé sur les classifications uniquement (frame matching).

pour chacune de ces configurations et déterminons leur efficacité dans l'évaluation extrinsèque, décrite en Section F.3.2.2 (attribution de rôles thématiques). Nous commençons donc par présenter l'évaluation au niveau de l'interface syntaxe/sémantique. La configuration la plus performante est ensuite utilisée dans l'évaluation plus globale des associations des verbes à des cadres syntaxiques et à des rôles thématiques.

3.4.1 Interface syntaxe/sémantique.

Comme le montre le Tableau 11, les associations entre clusters, cadres syntaxiques et rôles thématiques les plus adaptées sont obtenues en utilisant la configuration *Fpos* pour les cadres syntaxiques et *theta-2* pour les rôles thématiques. Pour les expériences suivantes nous utilisons donc les techniques d'étiquetage propres au clustering pour l'attribution de cadres syntaxiques, alors que les associations aux rôles thématiques sont obtenues par un alignement de groupes de verbes. Basé sur ces associations nous construisons nos étiqueteurs en rôles thématiques comme décrit en Section F.3.2.2 : l'un s'appuie uniquement sur la classification (frame matching) et l'autre combine la classification avec des estimations de probabilité. Comme pour FCA, nous évaluons les performances des étiqueteurs en rôles thématiques, basées sur ces associations, en fonction de plusieurs éléments : la granularité des représentations syntaxiques et la taille du corpus cible. Le clustering généré par IGNGF étant non-recouvrant, il n'y aura pas de différence entre le frame matching plus ou moins restrictif. La comparaison des étiquetages produits par ces systèmes aux annotations de SRL gold est résumé au Tableau 12. Le tableau montre que la performance est peu influencée par le type de représentation syntaxique utilisé : le système basé sur le frame matching atteint une F-mesure de 57 points, un tiers des instances ne sont pas étiquetées. Cette F-mesure est en dessous de celle du système par défaut, mais la précision est autour de 72%. Comme pour FCA, la F-mesure est améliorée pour l'étiquetage avec le système combinant frame matching et données probabilistes

(a) Résultats en utilisant une représentation détaillée des arguments syntaxiques (fonctions et plusieurs catégories phrastiques).

	BL	FAm	FAm+prob	FAm+p7+prob
F	65.21	57.05	62.82	62.81
Préc.	65.21	72.53	63.96	63.94
non étiqu.	0	35.17	3.50	3.47

(b) Résultats en utilisant une représentation moins détaillée des arguments syntaxiques (fonctions, catégories nominales et phrastiques uniquement).

	BL	FAm	FAm+prob	FAm+p7+prob
F	65.21	57.39	63.22	63.21
Préc.	65.21	72.69	64.07	64.04
non étiqu.	0	34.79	2.64	2.61

(c) Résultats en utilisant une représentation simple des arguments syntaxiques (fonctions seulement).

	BL	FAm	FAm+prob	FAm+p7+prob
F	65.21	57.16	63.61	63.39
Préc.	65.21	71.91	64.45	64.22
non étiqu.	0	34.04	2.55	2.55

(d) Résultats (approximatifs) de [Swier and Stevenson, 2005].

	BL	FAm	FAm+prob
F	74	76	83
non étiqu.	0	38% (16 amb.)	

Tableau 12: Résultats de l'étiquetage obtenu : en étiquetant les paires ⟨verbe, argument syntaxique⟩ de SRL gold seulement (*FAm+prob*) ou toutes les paires du corpus p7 (*FAm+p7+prob*). La représentation des arguments syntaxiques et plus (a) ou moins ((b) et (c)) détaillée. Les associations ⟨verbe, argument syntaxique, rôle thématique⟩ sont comparées à celles du SRL gold. Le Tableau 10d montre à titre indicatif les résultats de [Swier and Stevenson, 2005].

(63%). Toutefois, la précision diminue légèrement en dessous de celle de l'étiquetage par défaut mais en revanche presque toutes les instances sont étiquetées. Le rôle de la taille du corpus semble négligeable.

3.4.2 Niveau global.

Le Tableau 13 montre une comparaison des étiquetages au niveau des associations des verbes à des cadres syntaxiques et à des grilles thématiques par rapport au SRL gold. La ligne *Fpos* montre que la proportion des cadres présents à la fois dans la référence, le lexique et le clustering, de 48.91, est très basse. Une raison pour ceci est que les cadres très fréquents (comme le cadre transitive) ont une Feature F-mesure

(a) Associations entre verbes et **cadres syntaxiques**.

cadres synt.	SRL gold	classif	SRL gold et classif	SRL gold, lex, pas classif	SRL gold pas lex	Rappel	Rappel sans manquants en lex
Fpos	316	1100	134	140	42	42.41	48.91
Fpos+trans	316	1149	163	111	42	51.58	59.49

(b) Associations entre verbes et **grilles thématiques**.

SRL gold	SRL gold et classif	Rappel
318	124	38.99

Tableau 13: Évaluation sur un niveau plus global des associations entre verbes et cadres syntaxiques (a) et verbes et grilles thématiques (b) engendrées par la classification IGNGF (*classif*). Les associations aux cadres syntaxiques sont comparées aux annotations de la référence (*SRL gold*) et au lexique (*lex*). Les associations aux grilles thématiques sont comparées aux associations de la référence seulement. Sont représentés des types, ie. le nombre de paires ⟨verbe, cadre⟩ ou ⟨verbe, grille thématique⟩ distinctes.

très basse (du fait du processus de normalisation) et peuvent ne pas être associés au cluster. En ajustant le processus d'étiquetage pour prendre en compte les cadres partagés par 70% des membres du cluster nous obtenons un gain en rappel de 10% (ligne *Fpos+trans*).

Dans d'autres cas qui font défauts, deux classes Verbnet ont été regroupés et les traits associés caractérisent en fait l'une ou l'autre de ces sous-classes. D'autres erreurs sont liées au problème de polysémie. IGNGF, étant une méthode non-recouvrante, va toujours attribuer les verbes à une seule classe et par conséquent les verbes polysémiques à plusieurs ensemble de rôles thématiques, ne pourront pas être classés correctement.

En ce qui concerne les associations aux grilles thématiques le Tableau 13b montre que seulement 124 des 318 paires ⟨verbe, grille thématiques⟩ présentes dans le corpus sont engendrées par la classification. Une raison probable est à nouveau l'absence d'un traitement adéquate pour les verbes polysémiques.

3.5 IGNGF vs. FCA

Nous déduisons de ces expériences que la classification FCA présente de meilleurs résultats en terme d'associations de verbes aux cadres syntaxiques et grilles thématiques (niveau global), alors que IGNGF s'est montré plus utile dans l'attribution de rôles thématiques aux arguments syntaxiques des verbes (interface syntaxe/sémantique).

Le Tableau 14a présente une comparaison des cadres présents dans la référence SRL gold et ceux engendrés par les deux classifications.

Nous voyons que la classification FCA génère un plus grand nombre des paires

(a) Distribution des paires ⟨verbe, cadres⟩ dans SRL gold, le lexique syntaxique et les classifications IGNGF et FCA.

cadres	SRL gold	classif	SRL gold & classif	SRL gold & lex ¬ classif	SRL gold ¬ lex	Rappel	Rappel sans manquant en lex
IGNGF	316	1149	163	111	42	51.58	59.59
FCA	316	16542	243	31	42	76.90	88.69

(b) Nombre de paires ⟨verbe, cadres⟩ correctes dans les classifications FCA et IGNGF.

SRL gold ∩ lex	IGNGF ∩ SRL gold	FCA ∩ SRL gold	FCA ∩ IGNGF ∩ SRL gold
274	163	243	147

Tableau 14: Distribution des paires ⟨verbe, cadre⟩ dans la référence et les classifications FCA et IGNGF.

(a) Distribution des paires ⟨verbe, grille thématique⟩ dans la référence et les classifications obtenue par IGNGF et FCA.

Grilles thém.	gold	gold & classif	R
IGNGF	318	153	48.11
FCA	318	280	88.05

(b) Nombre de paires ⟨verbe, grille thématique⟩ compatibles avec les classifications FCA et IGNGF.

gold	IGNGF ∩ gold	FCA ∩ gold	FCA ∩ IGNGF ∩ gold
318	153	203	149

Tableau 15: Distribution des paires ⟨verbe, grille thématique⟩ dans la référence et les classifications IGNGF et FCA

⟨verbe, cadre⟩ de la référence que la classification IGNGF. Le nombre de ces paires présentes dans le lexique mais pas produites par la classification est plus élevé pour IGNGF que pour FCA. Une analyse des cadres problématiques montre que, pour les deux classifications, ceux-ci comprennent souvent les arguments AOBJ:PP et ATB:XP. Les associations de groupes de verbes à ces cadres s'avèrent souvent incohérentes indiquant des erreurs éventuels dans le lexique ou le besoin de traits sémantiques plus spécifiques à ces arguments.

Le Tableau 15a montre pour les deux classifications, le nombre de paires ⟨verbe, grille thématique⟩ de la référence où la grille thématique est compatible à une grille thématique associée au verbe par la classification. Une grille thématique est considérée compatible, si son ensemble de rôles thématiques est un sous-ensemble d'une grille associée au même verbe par la classification.

Ces résultats vont dans le même sens que ceux pour les paires ⟨verbe, cadre⟩ : les associations produites par FCA correspondent mieux à celles de la référence. La plupart des paires ⟨verbe, grille thématique⟩ de la référence qui étaient incompatibles à FCA étaient également incompatible à IGNGF. En même temps, comme le

(a) Résultats pour les associations ⟨verbe, argument syntaxique, rôle thématique⟩ basées sur *frame matching*, d'une part pour le clustering IGNGF, et d'autre part pour la classification FCA .

	%total (R)	%étiquetés (P)	F	%non étiqu.
baseline	65.21	65.21	65.21	0.00
FCA (partial)	30.87	70.40	42.92	56.14
IGNGF (partial)	47.43	71.91	57.39	34.79

(b) Résultats des associations ⟨verbe, arguments syntaxiques, rôle thématique⟩ produites par un système SRL combinant *frame matching* et un modèle probabiliste, en utilisant d'une part le clustering IGNGF, d'autre part la classification FCA.

		%total (R)	%étiquetés (P)	F	%non étiqu.
baseline		65.21	65.21	65.21	0.00
FCA (complete)	FM1, prob	57.23	59.80	58.48	4.30
IGNGF (basic)	FM, prob	62.39	64.07	63.22	2.64

Tableau 16: Résumé des résultats d'étiquetage en rôle sémantiques en utilisant d'une part le clustering IGNGF, d'autre part la classification FCA. Le système SRL est basé sur les classifications uniquement (*frame matching*, (a)) ou combine le *frame matching* avec un modèle simple probabiliste (b). Pour le système SRL par *frame matching* l'étiquetage basée sur FCA et IGNGF étaient le plus précis avec une représentation moins détaillées des arguments syntaxiques (partial, deux catégories syntaxiques). Le système combiné était le plus performant sur une représentation des arguments syntaxiques très détaillées (compléte, catégories syntaxiques détaillées) pour FCA et sur la représentation plus simple des arguments syntaxiques (basique, fonctions seulement) pour IGNGF. Chaque classification est utilisée dans la configuration où elle a montré les meilleurs résultats. La référence "baseline" est donnée pour des associations par défaut.

montre le Tableau 15b, beaucoup des paires correctes produites par IGNGF (149 sur 153, 97.36%) sont également générées par FCA, ce qui indique la cohérence de ces associations.

Nous comparons maintenant les deux méthodes de classification par leur capacité à attribuer des rôles thématiques à des instances ⟨verbe, argument syntaxique⟩ (interface syntaxe/sémantique). Chaque classification est utilisée dans deux systèmes d'étiquetage en rôles thématiques, l'un basé uniquement sur la classification (frame matching), l'autre combine l'étiquetage issu du frame matching avec des données statistiques. Les résultats sont présentés au Tableau 16. Ces résultats montrent clairement l'efficacité supérieure de la classification IGNGF dans cette tâche, en terme de F-mesure et de nombre d'instances étiquetées. Pour les systèmes basés sur les classifications uniquement (frame matching), le système IGNGF atteint une F-mesure de 57.39 avec 34.79% d'instances non-étiquetées contre une F-mesure de 42.92 et 56.14% d'instances non-étiquetées pour FCA. Les deux systèmes combinés parviennent à une amélioration considérable de la F-mesure (63.22 pour IGNGF et 58.48 pour FCA) et du nombre des instances étiquetées (97.36% pour IGNGF et 95.70% pour FCA). Néanmoins, pour tous les systèmes, la performance restent en

dessous de la baseline (attribution de rôles thématiques par défaut). Si les systèmes utilisant uniquement les ressources lexicales affichent une précision correcte (autour de 70%, au dessus de la baseline), les systèmes combinés font monter la F-mesure au détriment de la précision, qui baisse au dessous de la baseline (5% pour FCA mais seulement 1% pour IGNGF). Malgré cette comparaison défavorable avec la baseline, nous pensons que globalement ces résultats montrent la cohérence des associations ⟨verbe, cadre⟩ et ⟨verbe, grille thématique⟩ engendrées par les classifications et leur utilité dans cette tâche d'annotation sémantique.

4 Conclusion

4.1 Contributions

La contribution la plus importante de cette thèse est la proposition d'une approche innovante pour l'acquisition automatique d'une classification syntactico-sémantique pour les verbes du français, à partir de ressources lexicales françaises et anglaises existantes. En utilisant cette approche nous regroupons plus de 4200 verbes français et associons les classes verbales résultantes à des cadres de sous-catégorisation acceptés par ces verbes membres et en même temps à des ensembles de rôles thématiques représentant les participants des événements décrits par ces verbes. Telle que présentée ici, cette approche est légèrement supervisée, mais elle peut aussi être appliquée à des données distributionelles (issues de corpus), ce qui la rend non-supervisée et utilisable pour d'autres langues, à condition qu'un corpus et un analyseur syntaxique soient disponibles.

Pour créer cette classification nous avons exploré et adapté deux techniques de classification qui n'ont pas encore été utilisées dans ce contexte : une méthode de classification symbolique appelée analyse formelle de concepts (FCA – Formal Context Analysis) et IGNGF (Incremental Neural Gas with Feature Maximisation), une méthode neuronale (probabiliste) de clustering.

Les groupes de verbes créés sont reliés à des ensembles de rôles thématiques par une approche nouvelle, basée sur la traduction des classes de verbes Verbnet anglaises. Le résultat est une ressource où des groupes de verbes français sont alignés aux ensembles de rôles thématiques des classes Verbnet anglaises.

Nous avons effectué une évaluation approfondie des classifications produites, dans un premier temps par rapport aux groupements de verbes et dans un second temps basée sur les associations induites, de verbes avec les cadres syntaxiques et rôles thématiques. Bien que les classifications obtenues ne soient pas parfaites, nos ex-

périences ont montré que les classes verbales obtenues sont cohérentes d'un point de vue syntaxique et sémantique et se sont montrées utiles pour l'étiquetage en rôles sémantiques.

4.2 Perspectives

Information sémantique. Ils s'est avéré pendant nos expériences que l'information sémantique communément désignée par "restrictions" ou "préférences sélectionnelles" (selectional restrictions or preferences) a joué un rôle important et dans l'acquisition des classes verbales et dans leurs associations à des rôles thématiques. Une question intéressante est de savoir comment extraire et représenter cette information des ressources existantes et comment l'utiliser pour la constitution des classes syntactico-sémantiques.

[Mouton, 2010] propose une ressource française où les mots sont regroupés non seulement par champs lexicaux mais également par une sorte de similarité syntaxique. Ceci est une représentation possible de préférences lexicales et de ce fait il serait intéressant d'explorer leur apport à la constitution de nos classes verbales.

Un autre type d'information sémantique utilisée sont les rôles thématiques. Dans ce travail ils influent sur la nature et la composition des classes Verbnet traduites, qui déterminent à leur tour l'association des groupes de verbes avec les rôles thématiques. Il serait intéressant de déterminer l'ensemble de rôles thématiques le plus adapté à cette tâche et d'analyser d'une manière plus globale quels rôles tentent à se combiner à quels autres et l'influence que cela peut avoir dans leur réalisation syntaxique.

L'association des classes de verbes avec des rôles thématiques dépend également de comment les classes traduites sont alignées à nos classifications. Comme nous l'avons vu, notre méthode d'alignement basique souvent ne donne pas des résultats satisfaisants et une méthode plus sophistiquée, inspirée par exemple des techniques d'alignement d'ontologies, pourrait se révéler plus appropriée.

Méthodes de classification. Si d'une manière générale chacune des deux approches que nous avons utilisées a des atouts et des inconvénients, les deux se sont montrées peu performantes en ce qui concerne les associations aux cadres syntaxiques. Nous avons cependant, pour chaque méthode, identifié des approches pour palier ce manque.

Une autre question liée aux méthodes de classifications est de savoir comment les ajuster pour mieux prendre en compte le phénomène de la polysémie. FCA, qui est une classification recouvrante, permet de représenter la polysémie. Cependant elle à

tendance à attribuer les verbes à trop de classes alors qu'avec IGNGF, en raison de son caractère non-recouvrant, un verbe se retrouve toujours dans une seule classe.

Une direction de recherche prometteuse serait d'analyser les points communs des deux méthodes et d'explorer comment les traits s'ayant montrés utiles pour l'une pourrait être utilisés pour l'autre.

Connaissances linguistiques. Comme nos classifications permettent de relier des classes de verbes français et leurs cadres syntaxiques à des rôles thématiques anglais, elles permettent aussi de comparer la réalisation syntaxique de ces rôles dans les deux langues. Un premier exemple est la pronominalisation, un phénomène beaucoup plus courant en français qu'en anglais, où le cadre syntaxique comprend un pronom réfléchi clitique. Nos classifications permettent de capturer ce comportement syntaxique. Ainsi, il serait intéressant d'identifier les rôles thématiques et cadres syntaxiques anglais réalisés en français par des pronominalisations et aussi de savoir comment et quels rôles thématiques sont réalisés en français par des pronoms réfléchis.

List of Figures

List of Tables

Chapter 1

Introduction

Contents

1.1 Motivation

One of the goals of natural language processing (NLP) is to provide mechanisms for machines to understand the meaning of a text. To a large extent, this meaning is conveyed by predicates, typically verbs, which syntactically combine with other words to express events and how participants are related to these events. Therefore detailed knowledge about verbs and their syntactic and semantic behaviour is an essential ingredient in such an endeavour. This knowledge is typically stored in verb lexicons.

> This thesis is mainly concerned with the construction and evaluation of such a verb lexicon for French.

Although, there have been many different approaches to the construction of verb lexicons ([Dorr, 1998; Pustejovsky, 1995]) a clear consensus on how to build verb lexicons that are useful for NLP applications has not yet been developed. Nevertheless, approaches that are based on verb classes associating verbs with their elicit syntactic and semantic information has proven appealing: On the practical side, verb classes permit capturing generalisation about verb behaviour thus reducing both the effort needed to construct the verb lexicon and the likelihood that errors are introduced when adding new entries. On the theoretical side, [Levin, 1993] has shown

that the (syntactic) behaviour of a verb often reflects deep semantic regularities and that verbs belonging to a syntactic class often share a semantic meaning component. Lexical semantic classifications of verbs constructed in this way have been shown to support a wide range of NLP tasks, including lexical resource construction ([Korhonen, 2001]), natural language generation for machine translation ([Swift, 2005]), semantic role labeling ([Gildea and Jurafsky, 2002; Swier and Stevenson, 2004; 2005]) and information retrieval ([Klavans and Kan, 1998]).

> In this thesis we present an automatic acquisition method which permits constructing a classification of French verbs where the syntactic and semantic behaviour of verbs in a class is made explicit by associating the classes with syntactic frames (syntax) and thematic roles (semantics).

An important large scale verb classification for English which is built following the principles of [Levin, 1993] is Verbnet [Schuler, 2006]. For English, there exist several other large scale resources providing verb classes in a format that is amenable for use by natural language processing systems: FrameNet [Baker *et al.*, 1998], Verbnet [Schuler, 2006] and to a lesser extend Wordnet [Fellbaum, 1998].

> We chose to work with Verbnet in this thesis for the following reasons. Since it extends Levin's verb classes it is based on strong linguistic theoretical foundations and in the same time provides a large coverage. This makes it particularly useful for NLP applications. Furthermore, by working with Verbnet we hope to leverage the important amount of research which led to its construction, in the acquisition of a similar resource for a different language, namely French. Finally, one of our long term goals is to use the created resource in a semantic role labeling task for French. Because on one hand Verbnet thematic roles are less language dependent than other semantic role inventories and on the other hand have proven helpful in English semantic role labeling tasks, Verbnet seems to provide the kind of information which could eventually be used for semantic role labeling for French.

Only relatively few studies have been conducted on Levin style classifications for languages other than English ([Sun *et al.*, 2010; Brew and Schulte im Walde, 2002; Schulte im Walde, 2003; 2006; Oishi and Matsumoto, 1997; Dang *et al.*, 1998; Merlo *et al.*, 2002]), mostly focused on building verb classes using features extracted from distributional data acquired from corpora.

> Since our aim is to acquire a classification which covers the core verbs of French, we chose to use features extracted from manually validated

> lexical resources, rather than from distributional corpus data. Although these resources have less extensive coverage then corpora, they have often been built by human experts over several years and therefore contain valuable knowledge which is of a different nature and complementary to the distributional information available in corpora.

Most studies concentrate on acquiring sets of verbs which are semantically and/or syntactically coherent. The specific features characterising that coherence are usually left implicit: they determine the clustering of similar verbs into verb classes but they do not explicitly label these classes.

> In our approach we focus on explicitly associating groups of verbs with syntactic frames and thematic roles. While the associations with syntactic frames are inherent to the classification methods we use, the thematic roles are associated with the verb classes by using English Verbnet classes translated to French (drawing on the hypothesis that English Verbnet meaning components are valid across languages ([Jackendoff, 1990])).

We start from French and English lexical resources and build Verbnet like verb classes using two substantially different classification methods: The first is based on a symbolic classification technique called Formal Concept Analysis (FCA) and the second is a probabilistic neural clustering method (IGNGF, Incremental Growing Neural Gas with Feature Maximisation).

We compare and evaluate the classifications obtained with these two methods in different ways. First they are evaluated based on the verbs they group together, compared to a gold standard from the literature. Using this gold standard, we assess these groups of verbs semantically, that is, we check to what extent the classifications assign the verbs to the same semantic classes as the gold standard.

To evaluate the syntactic coherence of the classes and to assess their support to predict the syntax/semantics link between syntactic arguments and thematic roles, we evaluate the associations of verbs and frames and verbs and thematic roles induced by our classifications. For this, we created a second gold standard (called SRL gold), by manually annotating verbs and their syntactic arguments in a French treebank with thematic roles. This reference corpus allows us to assess the ⟨verb, syntactic frame, thematic roles⟩ associations engendered by our classification against those occuring in the manual annotations. Finally, we use the SRL gold reference to perform a task based evaluation. We assess the ability of our classifications to support the labeling of verbs and their syntactic arguments with Verbnet thematic roles by comparing the resulting labeling with the reference annotations.

The approach we propose for building Verbnet-like verb classes could be extended to other languages, provided the necessary resources are available. It could also be applied to corpus based data, thus making it fully unsupervised and directly applicable to any language for which a parser is available.

1.2 Road map

This thesis is organised as follows.

Chapter 2 describes previous work related to syntactic-semantic verb classes. We introduce and discuss various existing related resources and review methods applied for their acquisition.

Chapter 3 presents the resources and techniques we use for acquiring the French verb classes. These resources consist of existing lexical resources for French, the English Verbnet and three available translation dictionaries. Based on these resources we use two classification methods to build the verb classes. The first is a symbolic method, called Formal Concept Analysis (FCA) and the second a neural clustering method, called Incremental Growing Neural Gas with Feature Maximisation (IGNGF). We describe these techniques, show how they are applied to our data and the output classes they produce. We present first results and a preliminary qualitative evaluation.

The following chapters are devoted to a detailed evaluation of the verb classes produced with these two methods.

Chapter 4 is concerned with a semantic evaluation of the verb classes by comparing the generated groupings of verbs with an established gold standard which associates a sample of 170 French verbs to thematic role sets.

Chapter 5 consists of an evaluation of the obtained classifications based on the induced associations of verbs with syntactic frames and semantic role sets. For this we compare the acquired classifications with a reference corpus, where the syntactic arguments have manually been annotated with semantic roles.

Chapter 6 concludes by summarising the contributions of this thesis and presenting directions for further research.

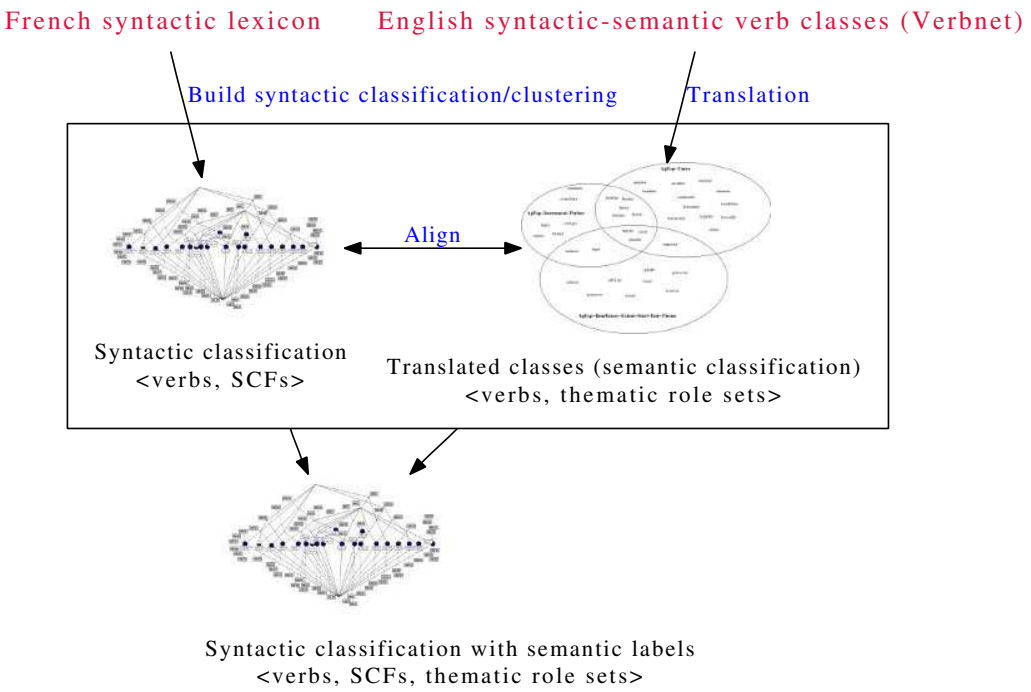


Figure 1.1: Outline of the procedure for creating syntactic semantic Verbnet-like classes for French verbs. Verbs are associated with syntactic features, namely SCFs, subcategorisation frames and semantic features, i.e. thematic role sets.

Since the creation of both classifications follows a similar pattern, we give a general overview of the procedure in Figure 1.1.

The left hand side of the picture shows how the associations of verb groups and syntactic frames are created: We start from a syntactic lexicon of French verbs and build a *syntactic* classification using a clustering/classification technique. Thus we produce groups of verbs which either are already associated with syntactic frames (in the case of FCA) or are labeled with these frames in a subsequent step (in the case of IGNGF). In both cases, the outcome is a *syntactic* classification where groups of verbs are associated with sets of subcategorisation frames.

On the other hand, as the right hand side of the picture shows, we start from the English Verbnet classes, translate them to French and obtain a *semantic* classification associating groups of French verbs with Verbnet classes and their thematic role sets. The *syntactic* and *semantic* classifications are then aligned (based on the member verbs of the classes), resulting in a syntactic-semantic classification which associates groups of verbs with both syntactic frames and sets of thematic roles. However, this picture is given for orientation only, in practice there is not always a clear-cut

difference between the *syntactic* and *semantic* classifications.

Chapter 2

Related Work on Semantic Verb Classes

Contents

In this chapter we review research relating to the topic of this thesis, namely the acquisition of syntactic-semantic verb classes for French. We first present English and French verb classifications where verbs are associated with syntactic and semantic

features. We then discuss various methods used for the automatic acquisition of such classes and finally show available evaluation strategies.

2.1 Verb Classes

Since in this thesis we are interested in verb classes where groups of verbs are associated with syntactic and semantic information, in this section we give an overview of existing related resources of this type. As noted in the introduction, for English, an important number of verb classifications amenable for use by natural language processing systems have been proposed. Here we are particularly interested in classifications which combine groups of verbs with syntactic and semantic features. In the following we first present some of these English classifications, focusing on the way classes are characterised in terms of syntactic and/or semantic features. We then briefly describe the only related French resources we are aware of, namely **Volem** and the **LADL** tables.

2.1.1 English Resources

2.1.1.1 FrameNet.

FrameNet ([Baker *et al.*, 1998]) is based on Fillmore's Frame Semantics ([Fillmore *et al.*, 2003]) and organises predicative lexical items (mostly verbs but also nouns and adjectives which can realise predicates) in so called "semantic frames" corresponding to situations which may be verbalised by the frame's members. A "semantic frame" is defined through fine grained semantic role labels, which are called Frame Elements. The lexical units (predicative items) are grouped solely on having the same frame semantics, a similar syntactic behaviour is not taken into account. Thus sets of verbs with similar syntactic behaviour may appear in multiple frames and a single FrameNet frame may contain sets of verbs with related senses but different subcategorisation properties.

FrameNet is built manually and currently covers 10 000 lexical units (word senses) which are associated with 958 frames and 2500 distinct frame elements. As the large number of Frame Elements suggests, most semantic roles are specific to individual frames.

The FrameNet resource is tightly connected with annotations of 150 000 sentences of the British National Corpus (BNC). Each annotation represents a Frame Element realisation and as such consists of the Frame Element name, a grammatical function (eg. object) and a phrase type (eg. noun phrase, NP).

To sum up, in the case of FrameNet, lexical units, in particular verbs, are grouped into so called frames, which are characterised semantically by the Frame Elements (thematic roles) a frame can have. The syntactic information is reported via the BNC corpus annotations but is not used to justify the membership of the predicates in the semantic classes.

2.1.1.2 Levin's verb classes.

Most of today's work on verb classes for Natural Language Processing (NLP) for English is strongly influenced by Beth Levin's seminal work ([Levin, 1993]). In her "Preliminary Investigation", Beth Levin provides a classification of English verbs which is guided by the hypothesis that there is a systematic relation between the syntactic and semantic properties of verbs. More specifically she defines verb classes based on the ability of each verb to occur or not in pairs of syntactic frames that are in some sense meaning preserving (diathesis alternations). In this way, Beth Levin first identified a set of 79 diathesis alternations and then manually classified about 3200 English verbs in about 200 classes, according to which of the diathesis alternation they entered. Although theoretically appealing, the verb classes as currently specified are still incomplete and not available electronically. In addition, Levin's classification does not explicitly provide semantic role labels, they are implicitly referred to by the "meaning preserving" alternations and thus difficult (impossible) to use in an NLP task. The original classes have been extended and made available on-line in Verbnet, which we discuss in the next section.

2.1.1.3 Verbnet.

Verbnet ([Schuler, 2006; Kipper *et al.*, 2008]) is inspired by and extends the Levin classification ([Levin, 1993], described in Section 2.1.1.2). It consists of hierarchically arranged verb classes. The Levin classification originally consisted of 240 classes organised into 47 top classes and 193 second and third level classes. Verbnet added almost 1000 lemmas to this classification and to each Levin class an explicit representation of the syntactic frames and semantic roles they express. The original Levin classes are extended by refining them whenever necessary to account for further semantic or syntactic differences within a class. Each class is characterised extensionally by its sets of verbs and, intensionally by a list of the arguments of those verbs and syntactic and semantic information about them. Table 2.1 shows a simplified example entry for the class *Hit-18.1.* With respect to the syntactic and semantic information we are interested in, the entry in this example gives the following

Class	*Hit-18.1*		
Parent	–		
Thematic roles	Agent, Patient, Instrument		
Selectional restrictions	Agent[+int_control] Patient[+concrete] Instrument[+concrete]		
Frames			
Name	**Example**	**Syntax**	**Semantics**
Basic Transitive	Paula hit the ball	Agent V Patient NP V NP	cause(Agent, E) manner(during(E), directedmotion, Agent) not(contact(during(E), Agent, Patient)) manner(end(E), forceful, Agent) contact(end(E), Agent, Patient)
Resultative	Paula kicked the door open	Agent V Patient Adj NP V NP ADJP	cause(Agent, E) manner(during(E), directedmotion, Agent) not(contact(during(E), Agent, Patient)) manner(end(E), forceful, Agent) contact(end(E), Agent, Patient) Pred(result(E), Agent, Patient)
Conative	Paul hit at the window	Agent V at Patient NP V PP	cause(Agent, E) manner(during(E), directedmotion, Agent) not(contact(during(E), Agent, Patient))

Table 2.1: Simplified VerbNet entry for the *Hit-18.1* class

information (for the *Hit-18.1* class):

Thematic roles The thematic roles participating in an event represented by the member verbs are Agent, Instrument and Patient.

Syntactic Frames The syntactic frames shown in this entry are: NP V NP; NP V NP ADJP; and NP V PP. A Verbnet entry not only gives the possible syntactic constructions the member verbs can occur in but also specifies linking, i.e. the surface realisation of the class's thematic roles in these constructions. Thus, in the example shown here the NP V NP construction is a syntactic realisation of the thematic grid Agent-Patient. This clearly allows to appropriately represent diathesis alternations:

Paula hit the ball	NP V NP	Agent V Patient
The ball hit the window	NP V NP	Instrument V Patient

In addition, applicable semantic roles are further characterised by selectional restrictions (eg. +concrete).

So, in contrast to FrameNet, the 30 Verbnet semantic roles are valid across classes. The link to syntactic structure is tighter for Verbnet than FrameNet: Classes are explicitly associated with syntactic frames, and in addition provide the linking of syntactic arguments and semantic roles (ie. they show how the semantic roles are realised as syntactic arguments of the subcategorisation frames).

Verbnet's role inventory has been chosen such that (i) verb arguments for all classes can be mapped to a semantic role and (ii) the roles appropriately convey key semantic components for each class. However, since there is evidence that a relatively small set of thematic roles will not cover all the possible arguments for all kinds of verbs, this set of roles is not claimed to be exhaustive, but the authors found that it offered enough descriptive informations for the 5200 verbs handled. These are the following (in parentheses the number of uses in classes): Actor (9), Actor1 (9), Actor2 (9), Agent (212), Asset (6), Attribute (4), Beneficiary (9), Cause (21), Destination (32), Experiencer (24), Extent (1), Instrument (25), Location (45), Material (6), Oblique (1), Patient (59), Patient1 (11), Patient2 (11), Predicate (23), Product (7), Proposition (11), Recipient (33), Source (34), Stimulus (5), Theme (162), Theme1 (13), Theme2 (13), Time (1), Topic (18), Value (5).

In contrast to FrameNet, there is no corpus annotated with Verbnet thematic roles.

2.1.1.4 PropBank.

PropBank ([Palmer *et al.*, 2005]) is not a verb classification and the primary goal in its development was not lexical resource creation but the development of an annotated corpus to be used as training data for supervised machine learning systems. The reason we present it here is that it adds a predicate-argument annotation to the syntactic structures of the Penn Treebank ([Marcus *et al.*, 1993]) and thus resulted in a lexical resource associating verbs and their syntactic arguments with thematic roles. PropBank consists of about one million words of the Wall Street Journal portion of the Penn Treebank II where syntactic arguments of verbs (nodes in the syntactic trees of the Penn Treebank) are assigned semantic role labels. This implies that the syntactic features associated to the verbs are determined by the Penn Treebank's syntactic representation. The semantic role labels were chosen to be generic and theory neutral but still consistent across syntactic variations of the same verb. Thus if *window* in *John broke the window* is annotated with the *Arg1* semantic role, it will be annotated with the same *Arg1* role in *The window broke.* However, because a universal set of semantic roles covering all types of predicates was unavailable, PropBank defined semantic roles on a verb by verb basis. While the *Arg0* and *Arg1* roles are in general prototypical Agents and Patients or Themes respectively ([Dowty, 1991]), for the higher numbered arguments no consistent generalisations can be made. Nevertheless, a set of roles corresponds to a distinct usage of a verb and is associated with a set of syntactic frames showing possible syntactic constructions realising this set of semantic roles for the given verb. There also have been efforts

to establish mappings between PropBank and Verbnet and FrameNet ([Loper *et al.*, 2007]) which show that often role assignments are consistent across Verbnet classes. In addition to the role labels *Arg0-Arg6* for the predicate's (verb's) complements which are considered arguments, PropBank defines more general roles which can apply to any verb and are similar to adjuncts. Some examples are *ArgLOC*ation, *ArgEXT*ent or *ArgTeMP*oral.

Due to its nature PropBank provides ample training data for semantic role labeling (SRL) systems, but it lacks information contained in Verbnet, as for example selectional restrictions, verb semantics and relationships between verbs. The linking between PropBank and Verbnet allows however to extend the machine learning techniques developed for PropBank to generate the more abstract Verbnet representations.

2.1.2 French Resources

Volem. Linguistically, Volem ([Fernandez *et al.*, 2002]) is closest to Beth Levin's verb classes. It is a resource built manually within the regional European project VOLEM and building on earlier work by Patrick Saint-Dizier ([Saint-Dizier, 1996; 1999]). The methodology is also very similar to that of Beth Levin: Patrick Saint-Dizier and his collaborators defined a set of diathesis alternations (called *contexts* in this framework) and assigned them to the verbs, according to whether the verb's usage in this *context* is acceptable or not. The initial set of *contexts* was enlarged and unified, in order to account for Romance languages other than French.

The resource is available from the authors as an xml file, a sample entry of which is shown below:

```
<TERME>
<VERBE>supprimer</VERBE>
<LCS />
<ROLETHEM>[[inic(agent)],[tid,tiv]]</ROLETHEM>
<ALTERNANCES>
   caus_2np,anti_pr_np,pas_etre_part_np_pp,
   pas_etre_part_np,state_2np_pp,caus_refl_pr_np
</ALTERNANCES>
<ALT-ANCIENNES>12,50,60,162,172</ALT-ANCIENNES>
<WN>8,2</WN>
<EXEMPLE>Ils ont supprimé ce mur</EXEMPLE>
</TERME>
```

Despite its affinity to Verbnet, the use of this resource is limited by its restricted coverage (1635 verbs) and non-standard representation format.

The LADL Grammar-Lexicon (aka LADL tables) [Gross, 1975; Guillet and Leclère, 1992; Boons *et al.*, 1976] is another manually built resource which gives detailed syntactic and semantic information about French verbs. It was developed by Maurice Gross and his collaborators at the "Laboratoire d'Automatique Documentaire et Linguistique" (LADL) from 1968 to 2002. Its initial purpose was to provide a systematic description of the syntactic properties of syntactic functors for French, in particular verbs. The LADL tables consists of a set of tables where each table groups together (usages of) predicative items that share some definitional properties. In particular, all predicative items in a given table share one (sometimes two) basic constructions (subcategorisation frames). For each predicative item present in the table, the columns of the table further specify the subcategorisation properties of that item. Typically, the table columns will provide: detailed information about the verb and about the possible realisations of its arguments (e.g., whether a given argument can be realised as a noun, as an infinitival or as a finite sentence; if an argument can be realised as a nominal, whether it may include a preposition and of what type; etc.), information reflecting syntactic properties of the verb or of the arguments (e.g., whether an argument can be cliticised), information about alternative subcategorisation frames and information about semi-regular redistributions true or not of a specific subcategorisation frame (e.g., whether a transitive syntactic frame admits the passive)[9].

These tables have been digitised by the Laboratoire d'Automatique Documentaire et Linguistique (LADL) and are now partially available[10] under an LGPL-LR licence[11]. Yet its use within natural language processing systems is still hampered both by its non standard encoding and by a structure that is partly implicit and partly underspecified.

2.1.3 Why Verbnet?

In this thesis we present an approach to the acquisition of French Verbnet like classes, based on the English Verbnet. So, a naturally arising question is: Why use an English syntactic-semantic resource for the description of French verb classes? The reason for this is that, since semantic roles are expected to be valid across languages, by using the same role inventory as for English, we hope to leverage some of the substantial research done for English and link syntactic information for French with semantic

[9] Examples are shown in Section 3.1.

[10] $\sim$ 60% of the table/verb information

[11] cf. `http://infolingu.univ-mlv.fr/DonneesLinguistiques/Lexiques-Grammaires/Presentation.html`

information provided by the English classes.

More specifically, we chose to aim at a Verbnet like classification for French verbs for the following reasons:

i) Verbnet semantic roles provide a compromise between generalisation and specificity in that, in contrast to FrameNet [Baker *et al.*, 1998] and PropBank [Palmer *et al.*, 2005] roles, they are common across all verbs, but are still able to capture specificities of particular classes.

ii) None of the other resources provides the link between syntactic arguments and semantic roles across different verbs (and classes).

iii) As shown in [Merlo and Van Der Plas, 2009], Verbnet provides a level of annotation which is less dependent on syntactic information and is therefore potentially more useful than PropBank in a cross-language application as ours.

iv) Using a role set already established for English allows to investigate similarities and differences in the syntactic realisation of these semantic roles in French.

2.2 Acquiring Verb Classes

The resources described in the previous section were all built manually. However, the manual creation of such resources requires a great effort: It takes time to define and acquire the lexical knowledge and, since language changes constantly the information needs to be updated regularly. Thus, such a lexicon will hardly ever be complete. For these reasons and also due to the increased algorithmic and computational capacities currently available, automatic induction of verb classes has become more accessible and a large variety of automatic or semi-automatic approaches have been proposed. Most of these approaches are based on supervised or unsupervised machine learning (ML) methods which draw the features needed for clustering from large corpora.

In the following we only give a very brief review of the most prominent techniques, for more detailed discussions please refer for example to [Schulte im Walde, 2009] and [Korhonen, 2009].

2.2.1 Machine learning methods.

Both supervised and unsupervised methods have been used. Supervised methods yielded good performance where the available training data was adequate and accurate. The employed techniques include K Nearest Neighbours, Maximum Entropy,

Support Vector Machines, Bayesian Multinomial Regression, Gaussian among others: [Joanis *et al.*, 2008] used Support Vector Machines, [Li and Brew, 2008] Bayesian Multinomial Regression whereas [Sun *et al.*, 2008] perform a comparative study of K Nearest Neighbours, Support Vector Machines, Maximum Entropy and Gaussian.

Unsupervised methods are particularly interesting for this work because they allow to learn classifications for languages or domains where no manually built classifications are available. The employed techniques include among others: K Means ([Schulte im Walde, 2006]), which is also often used as baseline, spectral clustering ([Sun and Korhonen, 2009]), Dirichlet Process Mixture Models ([Vlachos *et al.*, 2009]). In this work we exploit the clustering method IGNGF (Incremental Growing Neural Gas with Feature Maximisation, [Lamirel *et al.*, 2011b]) which to our knowledge has not yet been used for verb classification.

An important issue verb clustering methods have to address is verb polysemy: In language, verb polysemy is the rule rather than the exception, so assuming a crisp clustering, as is mostly done so far, is an over-simplification. A few approaches addressed this issue using soft-clustering algorithms and multiple assignment of verbs to clusters: [Pereira *et al.*, 1993; Rooth *et al.*, 1999] use an Expectation-Maximisation algorithm and [Korhonen *et al.*, 2003], Information Bottleneck, an iterative soft clustering method based on information-theoretic grounds. These experiments, in particular those presented in [Korhonen *et al.*, 2003], show that this is a hard problem in that polysemous verbs could not in general be assigned to classes corresponding to their different senses. Addressing this problem involves both providing a reference which appropriately captures the various cases of polysemy as well as finding a suitable clustering method. Yet, although polysemy clearly needs to be dealt with, so far there is no satisfactory approach in view for any of these issues.

All these approaches use features extracted from corpora. While most have focused on syntactic features in the form of shallow syntactic slots or verb subcategorisation frames, there also have been some attempts to refine the syntactic features with semantic information of verb selectional preferences: Sun and Korhonen could improve the verb clustering by using subcategorisation frames and selectional preferences as features.

However, the work addressed so far, mostly concentrates on acquiring verb classes which are semantically and/or syntactically coherent while the features characterising this coherence are usually left implicit: they determine the clustering of similar verbs into classes, but the classes are not explicitly labeled with syntactic and/or semantic attributes (in contrast to manually created classifications, in particular Verbnet).

The approaches are difficult to compare mainly because there is currently no

generally accepted scheme for evaluating verb classifications. One possibility is to use a gold standard. But currently available gold standards are small and only available for few languages. We will go more into detail discussing the evaluation issue in Section 2.3.

2.2.2 A Symbolic Method.

We know of only one symbolic automatic method employed to acquire verb classifications: Formal Concept Analysis (FCA), a methodology which we explore in this work. FCA [Barbut and Monjardet, 1970; Ganter and Wille, 1999], is a classification technique which permits creating, from a so-called formal context, a concept lattice where concepts associate sets of objects with sets of attributes. Here, the concept objects will be verbs while the attributes will be syntactic frames and semantic features. Intuitively, a concept is a pair $\langle O, A \rangle$ such that all the objects in O have exactly the attributes in A and vice versa, all attributes in A are true of exactly all the objects in O. That is, the concepts will group together sets of verbs which share exactly the same set of syntactic and semantic features. There are two major differences between classifications obtained with FCA and those resulting from one of the machine learning methods discussed above: First, FCA naturally produces a hierarchical and overlapping classification, while most clusterings obtained with numerical methods are crisp[12]. Second, whereas machine learning techniques assign a probability to a ⟨verb, feature⟩ association, with FCA the association is strict: If a verb and a feature are in the same concept, then the verb "had" this feature in the data. In addition, since the concepts associate exactly those verbs sharing the same set of features, FCA explicitly associates groups of verbs with subcategorisation frames and semantic features. We give a more detailed description of the FCA classification technique in Section 3.3.1.

2.2.3 Automatic Acquisition of Verb Classes for Languages Other than English.

While there has been much work on automatically acquiring verb classes for English, fewer studies have been conducted for other languages: [Schulte im Walde, 2006] presents approaches for German, [Merlo *et al.*, 2002] for Chinese and Italian, [Ferrer, 2004] for Spanish, [Oishi and Matsumoto, 1997] for Japanese and [Snider and Diab, 2006] for Arabic. In general, in these studies small or medium scale experiments

[12]There are however a few exception as for example the approaches presented in [Pereira *et al.*, 1993; Rooth *et al.*, 1999; Sun and Korhonen, 2011], which produce hierarchical and overlapping clusters.

are performed using various machine learning techniques: Thus [Schulte im Walde, 2006] uses K-Means to cluster 883 German verbs, [Merlo *et al.*, 2002] performs classification experiments on 59 italian verbs which are clustered using a decision tree algorithm, in [Ferrer, 2004], 514 Spanish verbs are clustered using a bottom up clustering algorithm. In all these experiments the clustering features are collected from corpus data and the resulting clusterings are evaluated either against manually built gold standards or by cross validation ([Merlo *et al.*, 2002]).

There are few studies on the automatic classification of French verbs. However, Sun *et al.*, apply a clustering approach developed for English (spectral clustering) to French and show that, as for English, syntactic frames and verb selectional preferences perform better than lexical cooccurence features. The exploited features are collected from a large scale subcategorisation lexicon acquired fully automatically from a French newspaper corpus. We will present two further approaches in this thesis: One is based on the use of Formal Concept Analysis and the other on the neural clustering technique IGNGF (Incremental Growing Neural Gas with Feature Maximisation). In contrast to the approach by [Sun *et al.*, 2010], the exploited features are extracted from French and English lexical resources.

So far, in this discussion we concentrated on building verb groupings. Indeed, most of the works presented earlier in this section, acquire verb classes which are coherent with respect to the exploited syntactic and/or semantic features. But as noted earlier, they don't explicitly name subcategorisation frames or thematic grids characterising a verb cluster (as is done in manually created resources, as for example Verbnet). Therefore, these classifications don't directly provide information associating verbs, syntactic frames and thematic role sets. On the other hand, a task aiming at this type of associations, namely of ⟨verb, syntactic argument⟩ pairs with thematic roles, is semantic role labeling (SRL). We may therefore hope to derive principles of mapping syntactic arguments and thematic roles from approaches adressing this task. In the following, we investigate how in selected SRL approaches ⟨verb, syntactic argument⟩ pairs are complemented with thematic role information. We concentrate on the methods relying on similar resources as the ones we dispose of.

2.2.4 Mapping Verbs and Syntactic Arguments to Thematic Roles

Mostly, relevant work in this respect make use of parallel texts and projection methods. That is, target ⟨verb, syntactic argument⟩ pairs are first aligned with source ⟨verb, syntactic argument⟩ pairs, based on the lexical semantic similarity of the target and source verbs. Then the thematic role associated to the source ⟨verb,

syntactic argument⟩ pair is projected to the target ⟨verb, syntactic argument⟩ pair. The aligned texts/sentences are of the same language or come from different languages. Considering first the approaches operating on the same language, [Fürstenau and Lapata, 2009], propose a method in which FrameNet annotations are extended by projection to not yet labeled sentences. [Lang and Lapata, 2010; 2011a; 2011b] induce the semantic roles of ⟨verb, syntactic argument⟩ pairs in the PropBank corpus using various clustering and graph partitioning approaches. Since the annotation scheme is that of PropBank, this system outputs verb-specific PropBank roles. Regarding now cross-language approaches, [Pado, 2007; Pado and Pitel, 2007; Padó and Lapata, 2009] assign FrameNet semantic roles by aligning nodes in syntactic constituent trees and projecting the semantic roles from English source annotations to German and French. Recently, [van der Plas *et al.*, 2011] proposed a semantic label transfer approach using a parallel English-French corpus (EuroParl). In this approach, the English part of the corpus is labeled with semantic roles using an SRL system trained on the English Penn Treebank corpus merged with PropBank labels. The semantic roles are then transferred to the aligned words in the French part. The French part of the corpus is parsed using the syntactic dependency parser described in [Titov and Handerson, 2010], trained on the dependency version of the French Paris 7 Treebank [Abeille *et al.*, 2003], which we also use in our experiments. Thus the transfer of the semantic roles results in effect in ⟨verb, syntactic argument, semantic role⟩ associations. These are then used to train a joint syntactic-semantic parser and to produce a semantic role labeler for French which outperforms the initial projected labeling and achieves results close to an upper bound from manual annotations.

Maybe the most relevant SRL approach to this work is the one proposed in [Swier and Stevenson, 2004; 2005]. It is an unsupervised bootstrapping method relying in the first place on the English Verbnet, which is used to produce probably unambiguous associations of ⟨verb, syntactic argument⟩ pairs with Verbnet thematic roles. These associations are then enhanced and refined using corpus data. Since we use this approach to evaluate our verb classifications, it is described in more detail in Section 5.2.1.2.

2.3 Evaluating Verb Classes

Currently there is no generally accepted scheme for automatically evaluating the induced verb classifications on the syntax-semantics interface. This is not too sur-

prising since, depending on the goal of the classification, different characteristics need to be assessed. The predominant question is of course whether the verb groups are coherent with respect to the classification criteria. But what is coherence concretely?

A characteristic trait of this type of classifications is that the groups of verbs are associated with syntactic and semantic attributes – how can we represent and adequately evaluate these associations?

Typically, in natural language processing, there are two kinds of evaluation methodologies: one by comparing against a reference or gold standard (*intrinsic evaluation*), and the second by checking the impact of the created resource in a natural language processing task (*extrinsic evaluation*). Both methodologies have been used for evaluating verb classifications. A third way of evaluating verb classifications, used for example in [Merlo and Stevenson, 2001], consists in cross-fold validation, where the annotated reference data is partitioned randomly and multiple experiments are then performed on one part of the data as input and the other part of the data as reference. The averaged results are compared to a baseline (typically a random attribution of verbs to classes) and an upper bound (typically inter-annotator agreement). We now briefly review the methods proposed in the literature to evaluate verb classifications.

2.3.1 Comparing to a Reference

For an intrinsic evaluation, a gold standard is needed. Most approaches refer to hand-crafted small-scale verb classes developed for the purpose of evaluation ([Sun and Korhonen, 2009; Sun *et al.*, 2010; Schulte im Walde, 2006; Merlo and Stevenson, 2001]). This is even more true for work in languages other than English. Other approaches make use of manually built resources, as Verbnet, or derive the gold standard from these: [Li and Brew, 2008; Joanis *et al.*, 2008; Sun and Korhonen, 2009; Ferrer, 2004]. For English however, two gold standards based on [Levin, 1993] have been used to evaluate several recent classifications ([Korhonen, 2009]). One has been introduced in [Joanis *et al.*, 2008] and consists of 835 verbs grouped into 15 Levin classes. The other is proposed by [Sun *et al.*, 2008] and classifies 204 medium-high frequency verbs into 17 fine-grained Levin classes. This gold standard is particularly interesting for our work because it has also been translated to French and can thus serve as a reference for this language as well ([Sun *et al.*, 2010]). These gold standards provide groups of verbs related to fine grained Levin classes. They do not explicitly give the syntactic and semantic features which (for Levin) characterise the classes. For English, these features can potentially be induced from the related Levin classes, this is however more problematic for other languages.

Even with a gold standard at hand, the evaluation task is not straight forward. For instance, it is not evident how the two sets of classes should be mapped onto each other, especially with different number of classes and it is not clear whether a given evaluation metric appropriately reflects the features which are most important for the sought classification. Finally, the choice of the evaluation metric is also influenced by comparability considerations: To relate to further work in the field one needs to employ a similar gold standard and a comparable or consistent evaluation metric.

[Schulte im Walde, 2003], Chapter 4, compared various evaluation methods against a gold standard. The following are considered most appropriate: (a) the f-score of a pair-wise precision and recall measure, (b) an adjusted pair-wise precision measure, and (c) the adjusted Rand index. In this work we use a measure to some extent similar to (a) (cf. Chapter 4 and 5).

In general, approaches are evaluated with various measures employed in clustering and classification research. Thus, [Joanis *et al.*, 2008] and [Li and Brew, 2008], whose verb classifications perform best according to [Korhonen, 2009], evaluate their approach using macro averaged recall. In other work ([Sun *et al.*, 2008; Sun and Korhonen, 2009; Sun *et al.*, 2010]) the measures used for evaluation are modified purity and weighted class accuracy (and their F-measure). Since [Sun *et al.*, 2010] proposed the only gold standard for French verb classes, we use their evaluation metrics in this work. These can be explained as follows. Each induced cluster is assigned the gold class (its *prevalent class*, prev(C)) to which most of its member verbs belong. A verb is then said to be correct if the gold associates it with the prevalent class of the cluster it is in. Given this, purity is the ratio between the number of correct gold verbs in the clustering and the total number of gold verbs in the clustering. Accuracy represents the proportion of gold verbs in those clusters which are associated to a gold class, compared to all the gold verbs in the clustering.

2.3.2 Task Based Evaluation

A task based evaluation consists in showing whether and to what extent the acquired classification can be used to support some natural language processing application. Although the work on automatic verb classification is largely motivated by their practical potential, we know of only few works where approaches to automatic verb classifications have been evaluated in this manner. In the following we review these approaches, together with the (few) other uses of syntactic-semantic verb classes in NLP applications we are aware of.

We identified the following applications involving syntactic-semantic verb classes. In [Li, 2008], Levin style verb classes, automatically acquired from corpus data, are

used to improve PP-attachment disambiguation. [Schulte im Walde *et al.*, 2008] build syntactic-semantic verb classes from the British National Corpus where the syntactic component is represented by subcategorisation frames, whereas the semantics are modelled by selectional preferences expressed as WordNet semantic concepts. The resulting verb classes are evaluated by using them to improve a language model.

Another area where verb classes are used is semantic role labeling: [Swier and Stevenson, 2004; 2005] use the syntactic frames and their associations with thematic roles provided by Verbnet to bootstrap a semantic role labeling system on sentences from the British National Corpus. We will use their method to evaluate the performance of our automatically acquired French Verbnet like classification to support a (simplified) similar semantic role labeling task.

[Giuglea and Moschitti, 2006] use the links between FrameNet, VerbNet and PropBank to build a semantic role labeling system on the FrameNet and PropBank corpora. They (semi-)automatically map FrameNet frames and Verbnet classes and thus interconnect FrameNet, VerbNet and PropBank with the effect of obtaining a better verb coverage and a more robust semantic parser.

This chapter provided a general overview of the problem setting for the task we are adressing in this thesis, namely the acquisition of a syntactic-semantic verb classification for French. We presented existing and available resources and motivated our choice to target a Verbnet like classification. We then reviewed automatic approaches for the acquisition of verb classifications and discussed existing evaluation schemes.

Chapter 3

Experiments

Contents

In this chapter we describe the experiments we performed for acquiring French Verbnet like classes. These experiments are based on two classification techniques: the symbolic method called Formal Concept Analysis (FCA) and the neural clustering method Incremental Growing Neural Gas with Feature Maximisation (IGNGF), which will be presented in Section 3.3. Since both methods group French verbs in existing lexical resources, based on features also extracted from these resources, we start by describing these resources and the extracted features in Section 3.1.

Our classifications are evaluated against two reference data sets, which are described in Section 3.2. Since these reference data sets influenced the choice of the data we performed our experiments on, they are presented together with the data sets we use in our experiments in this same section.

Finally, in Section 3.3 we describe the two classification/clustering methods we use. We give the theoretical background and show how they are adjusted to better fit the requirements of our classification task. We describe how they are applied to our data and show resulting sample classifications and preliminary evaluation results.

3.1 Lexical Resources and Feature Extraction

The syntactic and semantic features used for classification were extracted from existing lexical resources. In this section we present these lexical resources, describe the features used for classification and show how these features were extracted from the lexical resources.

3.1.1 Syntactic Resources

Syntactic features were retrieved from three existing lexicons for French namely, Dicovalence, TreeLex and the LADL tables. Recently, two other important syntactic lexicons for French have been made available, namely the Lefff [Sagot *et al.*, 2006] and LexSchem [Messiant *et al.*, 2008] lexicons. Both provide among other things subcategorisation information automatically extracted from large corpora. We chose however to build our classifications based on general-purpose lexical resources rather than on distributional data, since our aim was a classification covering the core verbs of French. We assumed that this task is better supported by general purpose

VAL$	expédier: P0 P1 (PL)
VTYPE$	predicator simple
VERB$	EXPEDIER/expédier
NUM$	41640
EG$	*la secrétaire a déjà expédié le courrier aux firmes intéressées*
TR_DU$	*verzenden, versturen*
TR_EN$	*send off, despatch*
FRAME$	subj:pron|n:[hum], obj:pron|n:[nhum,?abs], ?loc<>:pron|n:[]
P0$	qui, je, nous, elle, il, ils, on, celui-ci, ceux-ci
P1$	que, (qui), (te), (vous), la, le, les, (se réc.), en Q, ça, ceci, celui-ci, ceux-ci, (l'un l'autre)
PL$	0, où, y, là, ici, là-bas
RP$	passif être, se passif, (se faire passif)
AUX$	avoir

Figure 3.1: Dicovalence: sample entry for *expédier* with construction SUJ:NP,OBJ:NP (basic transitive construction).

lexica, whereas corpus based approaches are better suited for building and tuning verb classes tailored to specific corpus domains.

Dicovalence [van den Eynde and Mertens, 2003] is a manually constructed, syntactic lexicon for French verbs. It consists of more than 8000 entries listing the valency frames of 3936 French verbs. By valency frame is meant a specification of the number and nature of complements of a verb, including the subject, and indicating their syntactic function. In Dicovalence, this valency information is described according to the principles of the Pronominal Approach to syntax ([Eynde and Blanche-Benveniste, 1978], [Blanche-Benveniste *et al.*, 1984]). Following these principles, the dictionary specifies for each slot in a valency frame (syntactic argument) a description of the pronouns accepted in that slot. This description includes possible lexicalisations of that syntactic argument (for example whether the argument can be lexicalised as a noun or a phrase). It also specifies certain selectional restrictions and whether the frame can be used in various passive constructions. In Dicovalence, there is an entry for each verb and each valency frame this verb can be used with. Figure 3.1 shows the entry in Dicovalence for the verb *expédier* (*send*) when used with the basic transitive construction.

Thus, according to this entry, the verb *expédier* has the valency frame P0 P1 (PL) meaning that it can occur in a construction P0 *expédier* P1 PL, where P0, P1 and PL can be replaced by one of the entries described in fields P0$, P1$ and PL$ respectively. The brackets indicate that the PL$ argument is optional and may not be realised syntactically.

Dicovalence entries also contain a FRAME$ field which is a conversion of the

valency information into a format more suitable for NLP applications and more specifically in line with the currently used formalisms in French syntactic lexicons [Mertens, 2010]. We used the FRAME$ field to extract subcategorisation information. To make use of the information not provided by the FRAME$ field, we also extracted additional syntactic and semantic features from the Dicovalence entries, which were then used in the classification task. These syntactic features are the following:

Sym: The verb has a symmetric argument if the P0$, P1$ or P2$ field contains "se réc". This suggests that the verb can be used in a symmetric construction as in *L'un accuse l'autre des pires mensonges (One accuses the other of the worst lies.)/Ils s'accusent des pires mensonges. (They accuse themselves of the worst lies.).*

ArgNbr: The verb may have more than 4 syntactic arguments. This can be deduced from the FRAME$ field.

Event: One of the verb's arguments may be a phrasal construction. This can be derived from the FRAME$ field and suggests that one of the verb's argument may lexicalise an event.

Pred: The entry has a PX$ field. The PX$ field indicates a predicative complement as in *On l'a élu président./They elected him president.*

Theme: The verb has an optional object (an argument in parentheses in the valency frame) or accepts a passive reformulation formed using a reflexive pronoun (as in *ils se vendent (par eux)/they sell themselves (by themselves)*). According to [Randall, 2010], p. 95 and p. 120 respectively, this syntactic behaviour often indicates a Theme role.

The semantic features used in our experiments are the following:

Loc: The entry has a field PDL$ or PL$, which indicate the possible use of locative pronouns.

Nhum: The FRAME$ field contains a `nhum` indication (a non human argument).

Asset: The entry has a PQ$ field, which indicates a quantitative pronoun.

Plural: The FRAME$ field contains a `complex` indication which suggests a plural argument.

function	categories
SUJ	NP, Ssub, VPinf
OBJ	NP, Ssub, VPinf
AOBJ	PP, Ssub, VPinf
DEOBJ	PP, Ssub, VPinf
POBJ	PP, Ssub, VPinf
ATB	NP, PP, XP

Table 3.1: TreeLex syntactic function per syntactic category. Functions: SUJ (subject), OBJ (direct object), DEOBJ (indirect object introduced by de), AOBJ (indirect object introduced by a), POBJ (a complement with a different preposition), ATB (object's or subject's attribute)

We used these features because they possibly provide references to a thematic role or Verbnet class. This link is further discussed in Section 3.1.2.1. The Verbnet classes these features may help to predict are shown in Figure 3.5.

TreeLex [Kupść and Abeillé, 2008] lists subcategorisation frames for 2006 verbs. It was acquired automatically from the manually validated annotations of the Paris 7 Treebank [Abeille *et al.*, 2003] and verified manually. TreeLex subcategorisation frames list the syntactic arguments of the verb and give their function and category. Possible values for functions and categories are shown in Table 3.1. Thus, in the TreeLex syntactic representation, the basic transitive frame is represented as SUJ:NP,OBJ:NP.

The LADL tables have been introduced in Section 2.1.2. It is a manually built resource giving detailed syntactic and semantic information about French verbs. The data we could access contains 5076 French verbs grouped into 61 distinct tables whereby each table is associated with one or more defining subcategorisation frame. Figure 3.2 shows the entries for *expédier* in the LADL tables 38L and 3. The defining subcategorisation frames for tables 38L and 3 are the basic transitive frame with two nominal prepositional complements (SUJ:NP,OBJ:NP,POBJ:PP,POBJ:PP) and the basic transitive construction with a sentential prepositional object (SUJ:NP,-OBJ:NP,POBJ:VPinf) respectively, so, based on its occurrence in these 2 tables the verb was associated with these 2 frames. Since table descriptions give additional information which is not captured by the subcategorisation frames, additional syntactic and semantic information can be derived from a verb's membership in a table. We extracted the following syntactic features:

Sym: Symmetric arguments: Tables 36S, 36SL and 35S

(a) *expédier* in table 38L.

Verb	N0 source	N0 dest	N1 V	source seule	dest. seule	de	dans	sur	contre	à	...	example
expédier	-	-	-	+	+	+	-	-	-	+		*Max expédie les colis de Gap à Dax.*

(b) *expédier* in table 3.

Verb	N0 =: Nhum	N0 =: Nnr	N0 V	N0 V N1	N1 =: Nhum	N1 =: N-hum	N1 =: que P	N1 =: que Psubj	...	example
expédier	+	-	-	+	+	-	-	-		*Max expédie Léa chercher du vin.*

Figure 3.2: Ladl: sample entries for occurrences of verb *expédier* in tables 38L (a) and 3 (b).

ArgNbr: More than four syntactic arguments: Tables 18 and 38L

Pred: Table 39 groups verbs which can occur in constructions with two nominal arguments. In French this indicates a predicative complement (as for instance *président* in *On a nommé Paul président/One appointed Paul president*).

Event: The verbs allow sentential arguments: Tables 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16

The extracted semantic features are the following:

Loc: Location role: Tables 2, 3, 35L, 37E, 38LS, 38LH, 38LD, 38L, 35ST, 34L0, 37M4, 37M5, 37M6 and 38L1

Nhum: Concrete object (non human role): Tables 32C, 32A and 32CV

Asset: Table 32NM

Plural: Tables 32PL, 38PL, 36S, 35S and 36SL

Merging the 3 lexicons. We combined Treelex and the lexicons derived from Dicovalence and the LADL tables as follows: First, we used the TreeLex frame format to specify the frames assigned to the LADL tables. Second, we converted the Dicovalence frame representation into the TreeLex representation. This transformation was applied to the FRAME$ field value. Table 3.2 shows how the Dicovalence

syntactic functions and categories were converted to the TreeLex syntactic frame representations. For example the entry shown in Figure 3.1 was converted to the following entries in the merged lexicon:

expédier	SUJ:NP,OBJ:NP	*la secrétaire a déjà expédié le courrier* *the secretary has already sent the letter*
expédier	SUJ:NP,OBJ:NP,POBJ:PP	*la secrétaire a déjà expédié le courrier aux firmes intéressées* *the secretary has already sent the letter to interested firms*

In Appendix A we give the list of all frame representations used in our syntactic lexicon. The combined lexicon covers 5918 verbs, 345 subcategorisation frames and has a total of 20443 verb, frame pairs. It is publicly available at `http://talc.loria.fr/tl_dv2_ladl-a-subcategorisation.html`. Table 3.3 shows sample entries in this lexicon for the verb *expédier*. As can be seen, the merged lexicon specifies for each lexical entry, the name of the lexicon it was extracted from.

Extracted syntactic and semantic features. To sum up, these syntactic resources provide the subcategorisation information for roughly 6000 French verbs. In addition, the Dicovalence and LADL lexical resources provide the following syntactic features: **Sym** (the verb accepts symmetric arguments), **ArgNbr** (the verb has more than 4 arguments), **Event** (the verb has a sentential argument), **Pred** (the verb has a predicative argument), **Theme** (the verb has an optional object or accepts a passive reformulation using a reflexive pronoun). From Dicovalence and the LADL tables we also extracted the following semantic features: **Loc** (the verb has a Location thematic role), **Nhum** (one of its arguments is concrete), **Asset** (one of its roles may be an asset), **Plural** (one of its arguments implies a plural form or a form denoting a group).

3.1.2 Semantic Resources

We used the English Verbnet to obtain French semantic classes by translating the verbs in English Verbnet classes to French. The groups of French verbs in a translated Verbnet class were associated with the original English Verbnet class and its set of thematic roles. As shall be shown in Section 3.3.1 and 3.3.2, the translated classes were used in two ways: First they were aligned with verb clusters in our classifications in order to provide each cluster with a set of thematic roles. Second, we extracted semantic features from these classes which we then used to build the classifications.

Verbnet was described in detail in Section 2.1.1.3. Figure 3.3 gives a simplified representation of the *amuse-31.1* Verbnet class and highlights the components ex-

(a) DV → TreeLex functions.

DV functions	TreeLex		
subj	SUJ		
obj	OBJ		
objp	AOBJ	if prep = à	
	DEOBJ	if prep = de	
	POBJ	other prep	
attr	ATB		
attr_obj			
deloc	DEOBJ		
internal_cause			
loc	POBJ		
man			
objde	DEOBJ		
objà	AOBJ		
px	ATB	if no prep	eg. *nommer président*
	DEOBJ	if prep = de	eg. DV: *augmenter le café de prix*
quant	-	if no prep and OBJ already present.	
	OBJ	if no prep and no OBJ	
	DEOBJ	if prep = de	
	AOBJ	if prep = à	
	POBJ	if other prep	
eval	-		

(b) DV → TreeLex categories.

DV categories	TreeLex	
pron, n	NP	if function SUJ or OBJ.
	PP	if function POBJ, AOBJ or DEOBJ.
de_inf		
inf	VPinf	
à_inf		
compl		
de_ce_que_compl		
indirq	Ssub	
si_compl		
à_ce_que_compl		

Table 3.2: Conversion of Dicovalence frame format to TreeLex format: DV functions (a) and DV categories (b).

Verb: *expédier*	
SCF	**Source info**
SUJ:NP,DUMMY:REFL	DV:41640,41650
SUJ:NP,OBJ:NP	DV:41640,41650;TL
SUJ:NP,OBJ:NP,AOBJ:PP	TL
SUJ:NP,OBJ:NP,POBJ:PP	DV:41640
SUJ:NP,OBJ:NP,POBJ:PP,POBJ:PP	LA:38L
SUJ:NP,OBJ:NP,POBJ:VPinf	LA:3
SUJ:NP,POBJ:PP,DUMMY:REFL	DV:41640

Table 3.3: Sample entries in subcategorisation lexicon for verb *expédier*. The right column shows the lexicons the entry was extracted from: DV (Dicovalence), LA (LADL tables) or TL (Treelex).

verbs (242):	abash, affect, afflict, amuse, annoy, ...	
theta-grid:	EXPERIENCER [+animate] CAUSE	
frames (6):	NP V NP	EXPERIENCER V CAUSE
	NP V ADV-Middle	EXPERIENCER V Adv
	NP V NP-PRO-ARB	CAUSE V
	NP V NP PP.oblique	CAUSE V EXPERIENCER *with* OBLIQUE
	NP.cause V NP	CAUSE('s) OBLIQUE V EXPERIENCER
	NP V NP ADJ	CAUSE V EXPERIENCER Adj
	...	

Figure 3.3: Simplified Verbnet class *amuse-31.1.*

ploited to associate verbs with semantic features. This class groups 242 verbs; the thematic roles participating in events described by these verbs are Experiencer and Cause; and these roles are realised syntactically in the constructions shown in the **frames** field. We did not use the selectional restrictions (for example *+animate* in Figure 3.3) or the information represented by semantic predicates (not shown in Figure 3.3 but present in Verbnet).

3.1.2.1 Translating Verbnet classes.

To translate the verbs in the English Verbnet classes to French we used three translation dictionaries: Sci-Fran-Euradic, Google dictionary and Dicovalence [van den Eynde and Mertens, 2003]. Sci-Fran-Euradic is a French-English bilingual dictionary, built and improved by linguists, which contains 243539 pairs of French-English terms and is distributed by ELDA (`http://catalog.elra.info/product_info.php?products_id=666`). Sci-Fran-Euradic contains 40111 French-English verb pairs. From the Google dictionary (`http://www.google.com/dictionary`) we extracted 13824 French-English verb pairs. Finally, we derived 11351 verb translations from Dicovalence. The merged dictionary contains 51242 French-English verb pairs.

However, this translation is bound to be very noisy because verbs are polysemous[13] and the dictionaries typically give translations for several readings of the verb. Thus the dictionary may give several translations v_{fr} which do not correspond to the meaning given by the $\langle v_{\text{en}}, \text{class}\rangle$ pair or this meaning may even not be covered at all by the dictionary. We experimented with two methods to translate the English Verbnet classes, the first based on translation frequencies and the other using the Support Vector Machines (SVM) machine learning method.

[13]The average polysemy recorded by the Princeton WordNet for the various parts of speech is: 2.17 for verbs, 1.4 for adjectives, 1.25 for adverbs and 1.24 for nouns.

Translating Verbnet classes using translation frequencies. When translating the verbs in a given Verbnet class, we obtained a set of French verbs. Each of these French verbs might be the translation of several English verbs in the Verbnet class to be translated. For example, the *characterize-29.2* class has 48 English verbs: *adopt, bill, cast, certify, ...* and *adopt* has 6 possible translations, namely: *accepter, adopter, agréer, choisir, retenir, suivre*. Furthermore, *accepter* occurred as a translation of an English member of the *characterize-29.2* class 14 times, *agréer* 7 times and *suivre* twice. Based on these counts we considered that there is more evidence that *accepter* has a meaning corresponding to the *characterize-29.2* class than *suivre*. Following these heuristics, of the 335 French verbs which are translations of the English verbs in the *characterize-29.2* class we only kept those verbs whose translation frequency count was in the upper half of the counts for all the translated verbs of this class, i.e. we kept the $\lfloor 335/2 \rfloor = 167$ French verbs which were among the translations of an English verb of the *characterize-29.2* class more frequently. That is, we only kept in a translated Verbnet class those translations whose occurrence count for this class is in the upper half of the translation frequency distribution. In the following we will refer to this method as **median**.

When applying the **median** translation method to the 16 Verbnet classes in the gold standard which used for evaluation (cf. 3.2.1.1) we obtained the results shown in Table 3.4. This table shows that there were many more verbs in the classes of the translated Verbnet than in the English Verbnet.

We also explored another approach to create the translated classes, namely using the Support Vector Machines (SVM) machine learning method. This is described in the next section.

Translating Verbnet classes with Support Vector Machines (SVM). To translate English Verbnet classes using Support Vector Machines we reformulated the problem as a binary classification task, which is to decide whether a French verb is in a translated English Verbnet class or not. Thus the objects to be classified are pairs of French verbs and English Verbnet classes. To train the classifier, we first built a training set stating for each ⟨French verb, English VN class⟩ pair whether the verb could be said to have a reading belonging to the class. The classifier trained on this training set produced probability estimates for each ⟨French verb, English VN class⟩ pair, expressing the probability of the French verb of being a member of the English Verbnet class. Based on these probability estimates, we selected the most relevant pairs and the translated classes were obtained by assigning each verb in a selected pair to the corresponding English Verbnet class.

Verbnet class	fr-vn	en-vn
put-9.1	262	23
remove-10.1	133	40
send-11.1	116	26
get-13.5.1	215	34
hit-18.1	145	27
amalgamate-22.2	117	54
characterize-29.2	168	48
peer-30.3	47	20
amuse-31.1	476	242
correspond-36.1	83	36
manner_speaking-37.3	128	88
say-37.7	116	45
light_emission-43.1	36	22
other_cos-45.4	745	361
modes_of_being_with_motion-47.3	77	38
run-51.3.2	243	129

Table 3.4: Translated English Verbnet classes, number of French verbs kept (fr-vn), number of English verbs in class (en-vn).

The features. Each object to be classified, i.e. each ⟨French verb, English VN class⟩ pair was associated with a list of features. These features are listed in the following. They are all numeric and are similar to the scores used in [Mouton, 2010] to generate ⟨French lexical unit, FrameNet frame⟩ pairs.

Let v_{fr} be a French verb and C_{VN} be an English Verbnet class.

1. $|trans(v_{\text{fr}}) \in C_{\text{VN}}|$, number of English translations of v_{fr} which are in C_{VN}. This is the most obvious feature. Clearly, a high number of v_{fr}'s English translations is strong evidence that v_{fr} is semantically close to C_{VN}.

2. $\frac{|trans(v_{\text{fr}}) \in C_{\text{VN}}|}{|trans(v_{\text{fr}})|}$, percentage of translations of v_{fr} which are in C_{VN} compared to total number of translations of v_{fr}. This feature is a proportional version of Feature 1.

3. the number of English Verbnet classes containing a translation of v_{fr}.

4. the relative number of English Verbnet classes containing a translation of v_{fr}: Feature 3 / (total number of Verbnet classes).

5. $|C_{\text{VN}}|$, size of English Verbnet class.

6. the relative size of C_{VN}: $|C_{\text{VN}}|$ / (total number of verbs in Verbnet).

7. $|\{\langle v_{\text{fr}}, v_{\text{en}}\rangle | v_{\text{en}} \in C_{\text{VN}} \text{ with } (v_{\text{fr}}, v_{\text{en}}) \in \text{ dictionary }\}|$, number of translation pairs for the English verbs $v_{\text{en}} \in C_{\text{VN}}$.

8. the relative number of translation pairs:

$$\frac{|\{\langle v_{\text{fr}}, v_{\text{en}}\rangle | v_{\text{en}} \in C_{\text{VN}}, (v_{\text{fr}}, v_{\text{en}}) \in \text{ dictionary }\}|}{|\{\langle v_{\text{fr}}, v_{\text{en}}\rangle | v_{\text{en}} \in \text{ any VN class, } (v_{\text{fr}}, v_{\text{en}}) \in \text{ dictionary }\}|}$$

9. the number of French translations of C_{VN}: $|trans(C_{\text{VN}})|$.

10. the relative number of French translations of C_{VN}: $\frac{|trans(C_{\text{VN}})|}{|trans(\bigcup C_{\text{VN}})|}$

11. $|\bigcup_{C_{\text{VN}}} \{trans(C_{\text{VN}}) | v_{\text{fr}} \in trans(C_{\text{VN}})\}|$, the number of verbs in English Verbnet classes containing a translation of v_{fr}.

12. Feature 11, relative: Feature 11 / $|\bigcup(trans(C_{\text{VN}}))|$.

The training data. We produced the training data from the verbs in the gold standard described in Section 3.2.1.1. For each of these verbs v_{fr}, we collected $\langle v_{\text{fr}}$, Verbnet class $C_{\text{VN}}\rangle$ pairs as follows. v_{fr} was first translated to English giving a set of English verbs. These verbs are members of a set (Trans(v_{fr})) of English Verbnet classes: Trans(v_{fr}) = $\{C^1_{\text{VN}} \ldots C^k_{\text{VN}}\}$. Our training set then consisted of the pairs: $\{\langle v_{\text{fr}}, C_{\text{VN}}\rangle | C_{\text{VN}} \in \text{Trans}(v_{\text{fr}})\}$. Each of these pairs $\langle v_{\text{fr}}, C_{\text{VN}}\rangle$ was then manually annotated to indicate whether or not the association was correct, i.e. whether a reading of v_{fr} had the theta roles of C_{VN}.

As several Verbnet classes may have the same theta-grid and as it was not possible to decide whether a reading of a verb should be assigned to a given Verbnet class or not, we merged classes with identical theta-grid.

In addition, the discussion in Section 2.1 pointed out an observation holding for all semantic role labeling schemes: certain role labels seem to be more similar than others. However, currently there is no ultimate consensus about consistent groupings. In order to obtain groupings as consistent and coherent as possible for our classification experiments we proceeded as follows to group Verbnet thematic roles. In Verbnet's theta grids we replace Actor, Actor1, Actor2 by AgentSym, Patient, Patient1, Patient2 by PatientSym and Theme, Theme1, Theme2 by ThemeSym.

The Verbnet thematic roles were further grouped such that the number of resulting classes was as small as possible while still ensuring that roles were properly differentiated (i.e. in each Verbnet class all used thematic roles are distinct). Since we based our clustering experiments first on a restricted number of classes and second on all Verbnet classes, roles were grouped in two ways, as shown in Table 3.5. In the following, the first role set will be called **vn_restricted** and the second **vn_all**.

(a) Verbnet role groups for experiments based on a restricted number of Verbnet classes (**vn_restricted**).

Thematic role	VerbNet thematic roles
AgExp	Agent, Experiencer.
AgentSym	Actor, Actor1, Actor2
Theme	Theme, Topic, Stimulus, Proposition, Oblique
PredAtt	Predicate, Attribute
ThemeSym	Theme, Theme1, Theme2
Patient	Patient
PatientSym	Patient, Patient1, Patient2
Start	Material (transformation), Source (motion, transfer)
End	Product (transformation), Destination (motion), Recipient (transfer)
Location	
Beneficiary	
Cause	
Instrument	

(b) Verbnet role groups for experiments based on all Verbnet classes (**vn_all**).

Thematic role	VerbNet thematic roles
AgentSym	Actor1, Actor2, Actor
Agent	Agent
Experiencer	Experiencer
Patient	Patient
PatientSym	Patient1, Patient2
Theme	Stimulus, Theme
ThemeSym	Theme1, Theme2
Topic	Proposition, Topic
PredAtt	Attribute, Predicate
Start	Material, Source
End	Destination, Product, Recipient
Location	Location
Beneficiary	Beneficiary
Cause	Cause
Extent	Asset, Extent, Time, Value
Instrument	Instrument

Table 3.5: Verbnet role groups in clustering experiments: When using a restricted set of Verbnet classes (a) and when using all Verbnet classes (b).

We finally obtained the classes used in our experiments by merging the Verbnet classes as follows. We first replaced each role of a class by the group it has been assigned to and then grouped the verbs according to their theta-grid, which is given by the Verbnet class they belong to.

Returning to the training set, it consisted of pairs of verbs and theta grids with assessments of whether a reading of the verb may be considered to have the roles given by the theta-grid. We added to this list the verb, theta-grid pairs given by the gold standard. The resulting training set had different dimensions, depending on which role set was used to group the verbs. They are shown in Table 3.6. Table 3.7

Role set	# verbs	# Verbnet classes	# (verb, class) pairs
vn_restricted (3.5a)	161	70	1740
vn_all (3.5b)	161	77	1802

Table 3.6: Dimensions of training set used for SVM based translation. These dimensions depend on which role set is used to group Verbnet classes.

(a) Verbnet classes in gold, when using a restricted set of Verbnet classes (vn_restricted, Table 3.5a)

AgExp, Beneficiary, Extent, Start, Theme
AgExp, Cause
AgExp, End, Start, Theme
AgExp, End, Theme
AgExp, Instrument, Patient
AgExp, Location, Theme
AgExp, PatientSym
AgExp, PredAtt, Theme
AgExp, Start, Theme
AgExp, Theme
AgentSym, Theme

(b) Verbnet classes in gold, when using all Verbnet classes (vn_all, Table 3.5b)

Agent, Beneficiary, Extent, Start, Theme
Agent, End, Start, Theme
Agent, End, Theme
Agent, End, Topic
Agent, Instrument, Patient
Agent, Location, Theme
Agent, PatientSym
Agent, PredAtt, Theme
Agent, Start, Theme
AgentSym, Theme
Cause, Experiencer
Experiencer, Theme

Table 3.7: Classes occuring in the gold standard (cf. Section 3.2.1.1) obtained after grouping Verbnet semantic roles using only a restricted set of Verbnet classes (a) and all Verbnet classes (b).

shows the classes used in the gold standard (see Section 3.2.1.1) for the two role sets.

Building the translated classes. In this section we present how we used libsvm[14], an implementation of a support vector machine classifier on the training data obtained previously to build the translated Verbnet style classes.

We followed the procedure outlined in the practical guide available on the libsvm site. Of the training set produced in the previous section we randomly select 100 instances which we separated and held out as test set. The data is scaled and we performed a grid search using five fold cross-validation on the development set to determine the parameters. In the following we briefly introduce the parameters which need to be determined ([Chang and Lin, 2011]):

The goal of SVM is to produce a model based on the training data which predicts the target values of the test data, when given only the test data features. Given a training set of instance-label pairs $(\boldsymbol{x}_i, y_i),\ i = 1, \ldots, l$ where $\boldsymbol{x}_i \in \mathbb{R}^n$ and $\boldsymbol{y} \in \{1, -1\}^l$, the support vector machines require the solution of the following optimisation problem: $\min_{\boldsymbol{w}, b, \boldsymbol{\zeta}} \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{i=1}^{l}\zeta_i$, subject to $y_i(\boldsymbol{w}^T\phi(\boldsymbol{x}_i) + b) \geq 1 - \zeta_i, \zeta_i \geq 0$.

[14] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

Role set	Acc. on held out test set	Acc. at cross-validation parameter search	# pos. (verb, class) pairs
vn_restricted (Table 3.5a)	90.00%	93.84%	5164
vn_all (Table 3.5b)	83.00%	85.50%	5294

Table 3.8: Accuracy of SVM classification and number of (verb, class) pairs labeled with 1, when using restricted Verbnet classes and all Verbnet classes.

To be better separable, the training vectors $\boldsymbol{x}_i$ are mapped into a higher dimensional space by the function ϕ and the SVM then finds a linear separating hyperplane with the maximal margin in this higher dimensional space. $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = \phi(\boldsymbol{x}_i)^T\phi(\boldsymbol{x}_j)$ is called the kernel function. Here we used a radial basis function (RBF) as kernel: $K(\boldsymbol{x}_i, \boldsymbol{x}_j) = exp(-\gamma\|\boldsymbol{x}_i - \boldsymbol{x}_j\|^2), \gamma > 0$. Thus, the parameters to be determined are: $C > 0$, the penalty parameter of the error term and γ, the parameter in the kernel function.

Table 3.8 shows the accuracy on the held out test set compared to the maximum accuracy at cross-validation in the parameter search phase. We see that the performance is lower when using all Verbnet classes. While this should not come as a surprise since the training set only contains verbs in the restricted set of Verbnet classes, it shows that, when scaling to all Verbnet classes, the training set needs to be completed with training instances involving a larger number of Verbnet classes (ideally all). The parameters determined in the parameter search phase were then used to compute the probability estimates for the whole data-set of ⟨verb, class⟩ pairs. As can be seen in Table 3.8, 5164 respectively 5294 verb class pairs were labeled with 1 by the classifier. From this data we produced three sets of translated classes:

1. **svm** We selected 6000 ⟨verb, class⟩ pairs with highest probability estimates. We chose this number considering that Verbnet contains 5726 ⟨verb, class⟩ pairs.

2. **svm-best** We selected for each verb, the ⟨verb, class⟩ pair with highest probability estimate.

3. **svm-median** We selected for each verb, those ⟨verb, class⟩ pairs with probability estimates above median (with respect to this verb only).

Figure 3.4 shows the distribution of the French verbs in these different class sets compared to the number of English verbs in the corresponding classes and the number of French verbs in the translated classes obtained by keeping only the verbs with translation frequency above median. We used the role set shown in Table 3.5a, which is derived from the restricted set of Verbnet classes. We only show here the distribution of verbs in the classes occuring in the gold standard. Figure 3.4 suggests

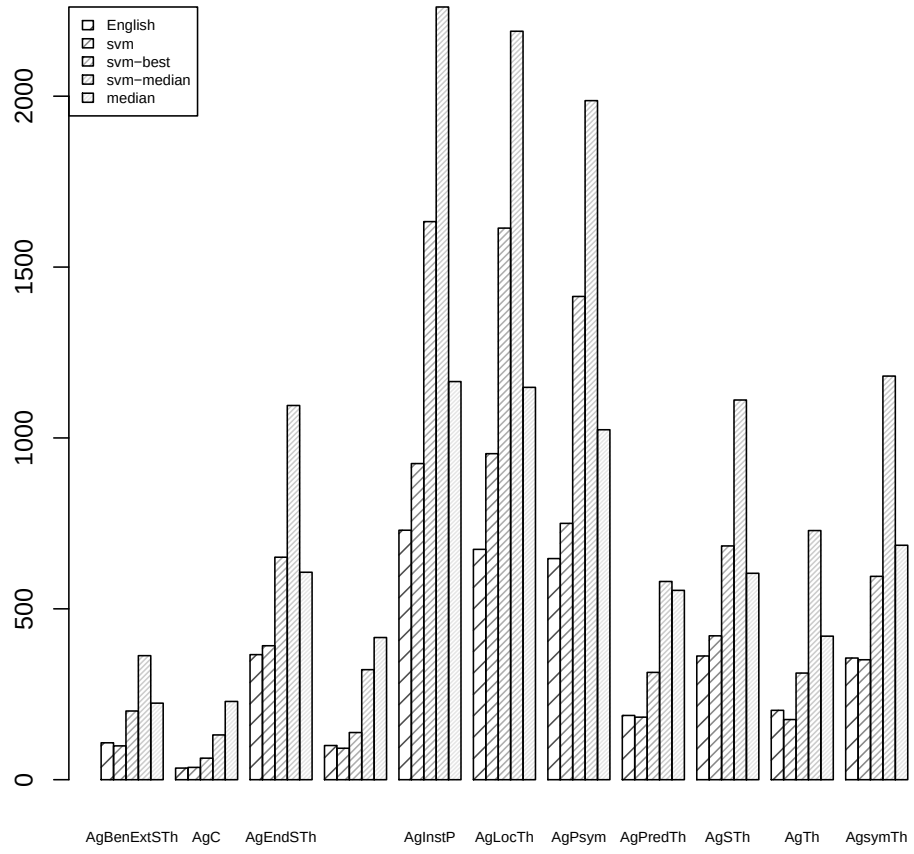


Figure 3.4: Distribution of verbs in classes with grouped theta-roles: English verbs in English Verbnet classes, French verbs in English Verbnet classes obtained in different ways by translation. The order of the bars in the bar groups is the same as in the legend.

that the translated classes closest to Verbnet with respect to the distribution of verbs is **svm**. In Section 3.3.1 we show how we used each of these class sets to associate groups of French with a thematic grid. The results of the experiments described there also suggest that the **svm** class set performed best.

We concluded from these experiments, that the translated classes produced with the **svm** method suited the requirements of our application best. Therefore in the following we use translated classes created with this method. To sum up, using the **svm** method, we created two sets of translated Verbnet classes: one called **vn_restricted**, which is based on a restricted number of Verbnet classes and one

called **vn_all** where we used all Verbnet classes. The Verbnet classes used in the **vn_restricted** set are those which contained a translation of a verb in the gold standard by [Sun *et al.*, 2010]. The selection of Verbnet classes used determined the groupings of thematic roles (shown in Table 3.5a for the restricted set and in Table 3.5b for all Verbnet classes) and by consequence the way Verbnet classes were merged (Table 3.7 shows the gold Verbnet classes with respect to the role groupings derived from the restricted class set (Table 3.7a) and all Verbnet classes (Table 3.7b)).

Extracted semantic features. The translated classes were used in the classification process as semantic features of verbs: a French verb is said to have the thematic grid θ as feature, if it is a member of a translated class whose corresponding English Verbnet class has the thematic grid θ. In addition, as we shall see in Section 3.3.1 and Chapter 5, the translated classes were used to characterise groups of French verbs. For this, groups of French verbs were aligned with the translated classes based on the verbs they have in common. A group of French verbs was then associated with the thematic grid of the "closest" translated class.

3.1.3 Feature Groups Used in Experiments

From the lexical resources presented above we extracted the following features which were used for clustering/classifying the verbs: the subcategorisation frames associated to each verb by the merged lexicon described in Section 3.1.1 and the thematic grid feature derived from the translated Verbnet classes as shown in Section 3.1.2. As explained in Section 3.1, we also extracted syntactic and semantic features from the Dicovalence and LADL resources. They are shown in Figure 3.5 together with Verbnet classes they can possibly help to predict.

These features were combined in several ways, resulting in eight feature sets which we use to build our classifications:

scf subcategorisation frame features only

scf-synt subcategorisation frames and syntactic features

scf-sem subcategorisation frames and semantic features

scf-synt-sem subcategorisation frames with both syntactic and semantic features

grid-scf subcategorisation frames and features derived from translated Verbnet classes

(a) Additional syntactic features.

Feature	**related VN class**
Sym, symmetric arguments	*amalgamate-22.2*, *correspond-36.1*
ArgNbr, 4 or more arguments	*get-13.5.1*, *send-11.1*
Predicate	*characterize-29.2*
Event type arguments (realised as clauses)	*correspond-36.1*, *characterize-29.2* ...
Theme, optional object	implicit theme [Randall, 2010], p. 95
Theme, Passive built with *se*	theme role [Randall, 2010], p. 120

(b) Additional semantic features.

Feature	**related VN class**
Location role	*put-9.1*, *remove-10.1*, ...
Nhum, concrete object (non human role)	*hit-18.1* (eg. INSTRUMENT) *other_ cos-45.4* ...
Asset role	*get-13.5.1*
Plural role	*amalgamate-22.2*, *correspond-36.1*

Figure 3.5: Additional syntactic (a) and semantic (b) features extracted from the LADL and Dicovalence resources and the alternations/roles they are possibly related to.

grid-scf-synt subcategorisation frames, syntactic features and translated Verbnet classes

grid-scf-sem subcategorisation frames, semantic features and translated Verbnet classes

grid-scf-synt-sem subcategorisation frames with both syntactic and semantic features and translated Verbnet classes.

3.2 Data Sets

In this section we describe the data sets we performed our experiments on and the reference data we used for evaluation. At first sight it may seem that these items do not belong into the same section. However, they are strongly related since, to be able to compare our results to the reference data, the verbs to be classified and their associated syntactic frames and thematic role sets had to be compatible with these references. Since the reference data influenced the data used in our experiments, we start by describing this reference data and then detail the data sets which we used to build our Verbnet like classifications of French verbs.

3.2.1 Evaluation Sets

For French there is only one evaluation resource we are aware of. This is the gold standard proposed by [Sun *et al.*, 2010] where about 160 French verbs are grouped into Levin classes. However, the usefulness of this gold standard is restricted because, since verbs are only associated to Levin classes, it does not permit an evaluation of the associations of the verb classes with syntactic frames. To evaluate the ⟨verb, frame, thematic grid⟩ associations produced by our classification we manually built a second gold standard (called SRL gold) where ⟨verb, syntactic argument⟩ corpus instances are labeled with Verbnet thematic roles. We first present the gold standard by [Sun *et al.*, 2010], called V-gold[15] in the following. We then present the corpus annotated with thematic roles, used to evaluate ⟨verb, syntactic frame, thematic grid⟩ associations.

3.2.1.1 V-gold: the gold standard by Sun *et al.*

As mentioned above, to evaluate our approach, we used the gold standard proposed for a Verbnet-style classification of French verbs by [Sun *et al.*, 2010]. This resource was derived from a similar English resource which originally consisted of 16 fine grained Levin classes with 12 (English) verbs each, whose predominant sense in English belong to that class. [Sun *et al.*, 2010] translated each member verb to French and all translations were considered. However candidate translations were only kept when they shared all diathesis alternations previously defined for that class using the criteria of [Levin, 1993]. As a result, 40% percent of the translations were discarded so that the resulting gold standard for French gathers 171 verbs in 16 classes. Classes are named according to the original Levin class. The classes contain 7 to 16 verbs and the average number of verbs per class is 10.7. The gold standard is shown in Table 3.9. The version by [Sun *et al.*, 2010] contained several verbs in their pronominal form: *s'associer, se promener, se déplacer*. However, after looking at the frames these verbs occurred with in our lexicon, we found that the pronominal versions were in fact alternating constructions of the non-pronominal forms and we therefore replaced the pronominal by the non-pronominal forms. It is however not clear whether in these cases the verbs should still belong to the same classes: Thus for example the non-pronominal form of *associer* is assigned to the *amalgamate-22.2* class which is characterised by the "symmetric patient alternation"[16]:

The merger associated company A with company B.	Agent V Patient1 Patient2
The merger associated the two companies.	Agent V Patient[plural]

[15] This standard is called V*erb*-gold, because it is used to evaluate groups of verbs.
[16] All English examples are taken from Verbnet.

On the other hand *s'associer* is assigned to the *correspond-36.1* class characterised by the "symmetric agent alternation":

I agreed with him about it.	Actor1 V Actor2 Theme
We agreed about it.	Actor[plural] V Theme

While *s'associer* does undergo the above alternation:

Je m'associe avec lui à ça.	Actor1 V Actor2 Theme
Nous nous associons à ça	Actor[plural] V Theme

associer arguably does not:

* Je associe avec lui à ça.	Actor1 V Actor2 Theme
* Nous associons à ça	Actor[plural] V Theme

Because of the small number of problematic cases we nevertheless kept the non-pronominal forms in the gold standard.

In our experiments we aligned groups of verbs with translated Verbnet classes. These Verbnet classes are based on two different thematic role inventories (see Section 3.1.2.1) and thus the verb groups were associated with sets of thematic roles from these role inventories. To be able to compare these alignments with the gold standard, we identify each Levin class in the gold with a thematic grid from the role inventory of the set of translated Verbnet classes used in the current experiment. These associations are made explicit in the first column of Table 3.9.

3.2.1.2 SRL gold

To evaluate the association between verbs, frames and grids provided by our classifications, we used a reference corpus (called SRL gold) where verb arguments were manually labeled with thematic roles. This reference corpus consists of sentences from the Paris 7 Dependency Treebank (P7, [Candito *et al.*, 2009]). It has been developed from a collection of newspaper articles from Le Monde, containing 350 931 tokens, 12 351 sentences and 25 877 verb instances. This collection was semi-automatically annotated with phrase structure trees [Abeille *et al.*, 2003] and manually verified, and the resulting treebank automatically converted to dependency structures [Candito *et al.*, 2009]. Since in the experiments where we compare our results with SRL gold we aimed at a large scale classification, they were performed on all the translated English Verbnet classes and on all verbs in our syntactic lexicons. Therefore, the role inventory used for both annotating the SRL gold instances and assigning thematic role sets was **vn_all** (Table 3.5b).

For the evaluation of our two verb classification methods, we annotated ⟨verb, syntactic argument⟩ instances in P7 which were chosen as follows. We annotated instances of the 116 verbs occuring in the treebank and the gold standard proposed

VN class and its thematic roles, restricted VN set all VN classes	**French translations kept in gold**
amalgamate-22.2 AgExp, PatientSym *Agent, PatientSym*	incorporer; associer; réunir; mélanger; mêler; unir; assembler; combiner; lier; fusionner
amuse-31.1 Cause, AgExp *Cause, Experiencer*	abattre; accabler; briser; déprimer; consterner; anéantir; épuiser; exténuer; écraser; ennuyer; éreinter; inonder
characterize-29.2 AgExp, PredAtt, Theme *Agent, PredAtt, Theme*	appréhender; concevoir; considérer; décrire; définir; dépeindre; désigner; envisager; identifier; montrer; percevoir; représenter; ressentir
correspond-36.1 AgentSym, Theme *AgentSym, Theme*	coopérer; participer; collaborer; concourir; contribuer; associer
get-13.5.1 AgExp, Beneficiary, Extent, Start, Theme *Agent, Beneficiary, Extent, Start, Theme*	acheter; prendre; saisir; réserver; conserver; garder; préserver; maintenir; retenir; louer; affréter
hit-18.1 AgExp, Instrument, Patient *Agent, Instrument, Patient*	cogner; heurter; battre; frapper; fouetter; taper; rosser; brutaliser; éreinter; maltraiter; corriger
light_emission-43.1 AgExp, Location, Theme *Agent, Location, Theme*	briller; étinceler; flamboyer; luire; resplendir; pétiller; rutiler; rayonner; scintiller
manner_speaking-37.3 AgExp, End, Theme *Agent, End, Topic*	râler; gronder; crier; ronchonner; grogner; bougonner; maugréer; rouspéter; grommeler; larmoyer; gémir; geindre; hurler; gueuler; brailler; chuchoter
modes_of_being_with_motion-47.3 AgExp, Location, Theme *Agent, Location, Theme*	trembler; frémir; osciller; vaciller; vibrer; tressaillir; frissonner; palpiter; grésiller; trembloter; palpiter
other_cos-45.4 AgExp, Instrument, Patient *Agent, Instrument, Patient*	mélanger; fusionner; consolider; renforcer; fortifier; adoucir; polir; atténuer; tempérer; pétrir; façonner; former
peer-30.3 AgExp, Theme Experiencer, Theme	regarder; écouter; examiner; considérer; voir; scruter; dévisager
put-9.1 AgExp, End, Theme *Agent, End, Theme*	accrocher; déposer; mettre; placer; répartir; réintégrer; empiler; emporter; enfermer; insérer; installer
remove-10.1 AgExp, Start, Theme *Agent, Start, Theme*	ôter; enlever; retirer; supprimer; retrancher; débarasser; soustraire; décompter; éliminer
run-51.3.2 AgExp, Location, Theme *Agent, Location, Theme*	voyager; aller; errer; circuler; courir; bouger; naviguer; passer; promener; déplacer
say-37.7 AgExp, End, Theme *Agent, End, Topic*	dire; révéler; déclarer; signaler; indiquer; montrer; annoncer; répondre; affirmer; certifier; répliquer
send-11.1 AgExp, End, Start, Theme *Agent, End, Start, Theme*	envoyer; lancer; transmettre; adresser; porter; expédier; transporter; jeter; renvoyer; livrer

Table 3.9: French gold classes and their member verbs presented in [Sun *et al.*, 2010]. The second and third row of the *VN class* columns show the thematic role set of the class when only a restricted number of Verbnet classes are used (2nd row, underlined) or when all Verbnet classes are used (3rd row, in italics).

in [Sun *et al.*, 2010] and described in Section 3.2.1.1. We chose these verbs because this is the only gold standard we know of which assigns French verbs to Verbnet like classes. For each of these target verbs (116 verbs which in the gold are distributed over 12 Verbnet classes), we randomly selected up to 25 sentences containing each of these verbs. In these sentences we labeled each occurrence of a ⟨verb, syntactic argument⟩ pair for a verb in the gold with a thematic role from the role inventory of **vn_all**. Thus 3605 verb arguments were labeled and 1600 verb instances associated with a thematic grid. We labelled from 1 to 46 verb instances, with roughly 14 instances per verb on average. To label the ⟨verb, syntactic argument⟩ pairs we proceeded using the following guidelines.

1) If the verb was used with the meaning represented in the gold standard we chose a role from the role set of the English Verbnet class the verb is associated with in the gold standard.

2) Else, we searched for an appropriate translation which is also in an English Verbnet class, and retrieved the corresponding role set for the English Verbnet class. If one of the roles in this role set reflected the meaning of the ⟨verb, syntactic argument⟩ pair, we used it as a label for this pair. If none of the roles reflected this meaning we used another semantic role from Table 3.5b.

3) If we could find no appropriate translation which was also in an English Verbnet class, we used another semantic role from Table 3.5b.

For illustration we show some annotation decisions for syntactic arguments of the verb *révéler*, which in the gold is associated with the *say-37.7* class, which has the thematic grid Agent-End-Topic.

Utterance 1

Daniel Yergin	*révèle*	*la problématique.*	(*Daniel Yergin reveals the problem.*)
subject	V	object	
Agent	V	Topic	*say-37.7*

Here the syntactic arguments to be associated with thematic roles are the subject, *Daniel Yergin*, and the object *la problématique*. We considered that in this utterance "révéler" has the "say" sense and that the syntactic arguments subject and object could appropriately be labeled with the roles Agent and Topic respectively.

Utterance 2

Les choix ...	*révèlent*	*une ... indécision*	(*The choices reveal ... indecision*)
subject	V	object	
Cause	V	Topic	*indicate-78*

AOBJ:PP	AOBJ:VPinf	
ATB:XP		
DEOBJ:PP	DEOBJ:VPinf	
OBJ:NP	OBJ:VPinf	OBJ:Ssub
SUJ:NP		

Table 3.10: Syntactic arguments occuring in reference annotations.

Again, the syntactic arguments to be labeled are the subject (*Les choix*) and the object (*une indécision*). Here we considered that the verb *révéler* is not used with the *say-37.7* meaning but rather with the meaning of the English translation *reveal* as a member of the *indicate-78* class. This class has the thematic grid Cause-Recipient-Topic and the syntactic arguments of *révéler* (the subject *le choix* and the object *une indécision*) can be appropriately labeled with the roles Cause and Topic respectively.

Utterance 3

La stratégie	*se*	*révèle*	*payante*	*The strategy reveals itself as paying off*
subject	reflexive clitic	V	attribute	*The strategy proved as paying off*
Theme		V	PredAtt	*declare-29.4*

Here we found that the verb is used in the *declare-29.4* meaning with the thematic roles Agent-Theme-PredAtt. Using these roles we labeled the attribute syntactic argument *payante* with the PredAtt role and the subject (*la stratégie*) with the Theme role. With respect to *se*, we considered it an expression of the French "se passive" alternation and therefore did not associate it with a thematic role.

We thus ended up with 427 distinct ⟨verb, syntactic argument⟩ pairs (types) and 3605 instances of such pairs (tokens) which were labeled with the semantic roles in Table 3.5b. The syntactic arguments occuring in the reference are shown in Table 3.10.

The corpus instances were labeled with the following thematic roles (in parentheses the number of instances labeled with the corresponding role): Agent (1263), Beneficiary (25), Cause (27), End (302), Experiencer (59), Extent (2), Instrument (79), Location (48), Patient (261), PredAtt (141), Start (60), Theme (1106), Topic (237).

3.2.1.3 SRL gold vs. merged syntactic lexicon vs. V-gold.

We are now in a position to compare the associations of verbs with syntactic frames and thematic grids provided by our merged syntactic lexicon (Section 3.1.1) with those present in the V-gold (by [Sun *et al.*, 2010]) and the SRL gold. Recall that the

lexicon provides ⟨verb, syntactic frame⟩ associations, whereas the gold assigns verbs to Verbnet classes (and their thematic grids).

First, with respect to the lexicon, of the 316 *subcategorisation frames* (types) assigned to the verbs in the corpus, 42 types corresponding to 87 tokens were not present in the lexicon, suggesting possibly lacking entries (see Table A.2 in Appendix A).

Regarding the *thematic grids* and the V-gold, of the 1600 ⟨verb, thematic grid⟩ annotations, 78.60% were compatible with the Verbnet class assigned to that verb by the V-gold, that is, the thematic roles realised in the grids were a subset of the roles of the Verbnet class the verb was associated with by the gold. However, in the remaining 21.40% cases the verb was used in the utterance with a sense differing from the one given by the gold class. Per verb, the percentage of compatible thematic grids ranged from 0 to 100% (0 to 44 instances), with an average of 76.38% (roughly 11 instances).

Considering now the 427 distinct ⟨verb, syntactic argument⟩ pairs in SRL gold, 326 (76.34%) were associated with a thematic role which was present in the V-gold standard. Of the 3605 token associations, 2994 (81.18%) occurred in the gold standard. This shows that an important number of instances are occuring in the corpus with another sense than the one presumed in the gold standard.

3.2.2 Data Sets

We performed our experiments on two data sets. The first is based on a restricted set of Verbnet classes (**vn_restricted**), whereas for the second we used all Verbnet classes (**vn_all**). Each data set consists of a set of French verbs and the features associated to these verbs by our lexicons and a set of translated Verbnet classes and their sets of thematic roles (i.e. the roles of the English source classes).

The vn_restricted data set was derived from the V-gold standard, introduced in [Sun *et al.*, 2010] and described in Section 3.2.1.1. This gold standard consists of about 160 verbs grouped in Levin classes as shown in Table 3.9. The translated English Verbnet classes used in this data set are produced as described in Section 3.1.2.1. For this data set we used the set of translated classes called **vn_restricted** in Section 3.1.2.1. As described there, these were built using the **svm** method and only those Verbnet classes which contained a translation of a verb from the V-gold standard. These Verbnet classes are associated with sets of thematic roles from the role inventory shown in Table 3.5a. The verbs contained in this data set are the verbs occuring in our syntactic lexicon and the verbs in the **vn_restricted** set of translated

Data set	# verbs	# scfs	# thematic roles	# translated classes
vn_restricted	2091	238	13	70
vn_all	4260	303	16	77

Table 3.11: Dimensions of data sets used in clustering/classification experiments.

Verbnet classes (that is the translated Verbnet classes which contain at least one verb from the gold standard). Verbs with one subcategorisation frame only were ignored. Classifications based on this data set were evaluated against the gold standard by [Sun *et al.*, 2010], described in Section 3.2.1.1. Table 3.11 shows the dimensions of this data set.

The vn_all data set was built using all the verbs in our syntactic lexicon and all translated English Verbnet classes. Again, the set of translated English Verbnet classes was produced using the **svm** method as described in Section 3.1.2.1 starting from all English Verbnet classes and resulting in the set of classes called **vn_all** in Section 3.1.2.1. The role inventory used for these classes is shown in Table 3.5b. The verbs in this data set are the verbs occuring in the syntactic lexicon and the verbs in all translated Verbnet classes. Again, verbs with one subcategorisation frame only were ignored. Some counts for this data set are summarised in Table 3.11.

For both data sets, the translated Verbnet classes were obtained using the **svm** method (described in Section 3.1.2.1) because, as shown in Section 3.3.1.3, this method produced the ⟨verb, thematic grid⟩ associations most similar to those in the gold standard by [Sun *et al.*, 2010].

3.3 Clustering Methods

We applied two clustering/classification techniques to produce Verbnet like classifications. The first is a symbolic classification technique called Formal Concept Analysis (FCA) and the second is called IGNGF (Incremental Growing Neural Gas with Feature Maximisation) and is a probabilistic method based on a neural clustering procedure. FCA simultaneously groups the verbs by the syntactic frames they can be used with and vice versa the syntactic frames by the verbs accepting them, while IGNGF clusters the verbs based on their features provided by the lexicon. That is, to produce Verbnet like classifications, the FCA classes needed to be associated with thematic role sets, whereas for the IGNGF clusters we needed to identify a group of subcategorisation frames and a thematic role set characterising the verbs in the cluster. In the following we first briefly introduce the theoretical background

of each of these methods. We then show how they are applied to create groups of French verbs associated with subcategorisation frames and thematic grids. That is, in Section 3.3.1, for FCA we first show how we use this technique to produce concepts simultaneously grouping verbs and subcategorisation frames and then describe in Section 3.3.1.3 how these concepts are associated with thematic role sets. For IGNGF (Section 3.3.2) we first explain how the verbs are clustered based on the features extracted from the lexical resources and then in Section 3.3.2.4 show how the resulting clusters are assigned groups of subcategorisation frames and a thematic role set.

3.3.1 Formal Concept Analysis (FCA)

FCA [Barbut and Monjardet, 1970; Ganter and Wille, 1999] is a classification technique which permits creating, from a so-called formal context, a concept lattice where concepts associate sets of objects with sets of attributes. Here, the objects were the verbs and the attributes were syntactic frames.

Intuitively, a concept is a pair $\langle O, A\rangle$ such that all the objects in O have exactly the attributes in A and vice versa, all attributes in A are true of exactly all the objects in O. That is, our concepts group together sets of verbs which share exactly the same set of syntactic and semantic features (and vice versa, the features which are valid for for the same set of verbs).

More formally, a formal context $\mathcal{K}$ is a triple $\langle \mathcal{O}, \mathcal{A}, R\rangle$ such that $\mathcal{O}$ is a set of objects, $\mathcal{A}$ a set of attributes and R a relation on $\mathcal{O} \times \mathcal{A}$. Given such a context, a concept is a pair $\langle O, A\rangle$ such that

$$O = \{o \in \mathcal{O} \mid \forall a \in A.\ (o, a) \in R\}$$
$$\text{and vice versa}$$
$$A = \{a \in \mathcal{A} \mid \forall o \in O.\ (o, a) \in R\}.$$

Two operators, both denoted by $'$, connect the power sets of objects $2^{\mathcal{O}}$ and attributes $2^{\mathcal{A}}$ as follows:

$$' : 2^{\mathcal{O}} \rightarrow 2^{\mathcal{A}},$$
$$X' = \{a \in \mathcal{A} \mid \forall o \in X.\ (o, a) \in R\}$$

The operator $'$ is dually defined on attributes. For a formal concept $\langle O, A\rangle \in$

$O \times A$ we have $O' = A$ and $A' = O$. O is called the *extent* or *extension* and A the *intent* or *intension* of the formal concept.

A concept $C1 = \langle O1, A1 \rangle$ is smaller than another concept $C2 = \langle O2, A2 \rangle$ (written $C1 \leq C2$) iff $O1 \subseteq O2$ and $A1 \supseteq A2$.

The set of all formal concepts of a context $\mathcal{K}$ together with the order relation $\leq$ form a complete lattice called $\mathbb{K}$, the concept lattice of $\mathcal{K}$. That is, for each subset of concepts there is always a unique greatest common sub-concept and a unique least common super-concept.

3.3.1.1 Building the concept lattice based on French verbs and subcategorisation frames.

To associate French verbs with syntactic frames, we applied FCA on a formal context, where objects are French verbs and attributes are essentially the subcategorisation frames associated to these verbs by the merged lexicon described in Section 3.1[17]. Applying FCA to this formal context resulted in a concept lattice where the verb classes were the concept extents and the frame classes the concept intents. Thus, this concept lattice is in effect a verb classification, where verbs are grouped according to the subcategorisation frames they can be used with, and in the same time a classification of the syntactic frames, according to the verbs accepting them. The hierarchical structure of both, the verb and frame classes is given by the $\leq$ relation on the concepts.

For illustration consider the following simple example, derived from the introduction in [Levin, 1993], where Levin shows that *break, cut, hit* and *touch* verbs can be grouped according to their syntactic behaviour. This example is too small to demonstrate verb classification (more so as each of these verbs is from a separate Verbnet class) but it hopefully still gives an intuition of how FCA can be employed at the acquisition of verb classes. For the four verbs *break, cut, hit* and *touch* we extracted from the corresponding Verbnet classes[18] the syntactic frames these verbs can be used with[19] and obtained the formal context shown in Figure 3.6a. The resulting concept lattice is shown in Figure 3.6b. When read from top to bottom, the concept lattice can be interpreted as a classification of the verbs with respect to the frames, and from bottom to top as a classification of the frames with respect to the verbs. The arrows between the concepts represent the $\leq$ relation. $c_j \leq c_i$ implies that the

[17] Parts of this section are published as [Falk and Gardent, 2011]

[18] These are: *break-45.1, cut-21.1, hit-18.1* and *touch-20*.

[19] We only used the syntactic frames relevant for the Middle, Conative and Body-Part Possessor diatheses alternations, which Levin uses in her illustration.

(a) Formal context. Objects are verbs *break, cut, hit* and *touch*. Attributes are syntactic frames extracted from Verbnet.

	break	cut	hit	touch
NP V	×			
NP V ADV	×	×		
NP V NP	×	×	×	×
NP V NP PP	×	×	×	×
NP V PP		×	×	
NP V PP PP		×	×	

(b) Resulting concept lattice.

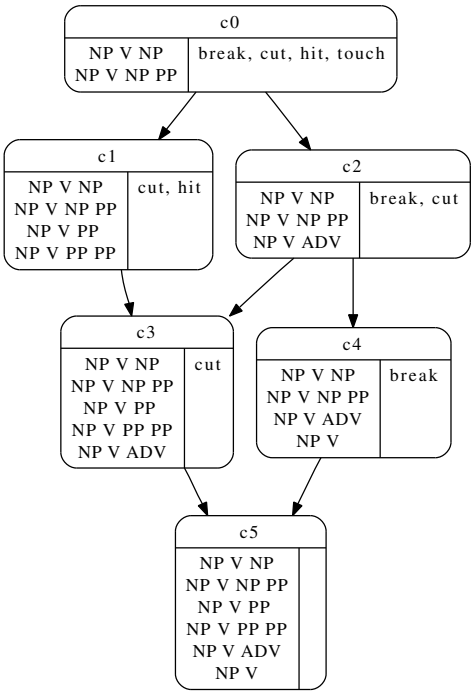


Figure 3.6: Applying FCA for the classification of verbs according to the syntactic frames they can be used in.

verbs in concept c_j are in concept c_i and the frames in c_i are in c_j. In this particular example, all verbs are transitive and may be used with an additional prepositional object, therefore the root concept c_0 has these two syntactic frames. In contrast, there is no verb which can be used with all syntactic frames (the bottom concept c_5 has no verbs). There are two characteristic frame sets for the verbs *break* and *cut* (these verbs are the only verbs appearing in the concepts c_4 and c_3 respectively), but the other two verbs could not be singled out in a similar way[20]. There are many ways this lattice can be turned into a verb classification. The most straight forward way, with minimum loss of information, is to remove the bottom concept, which typically contains no verbs. We could also only keep the three top concepts and thus obtain a more coarse grained classification still covering all verbs.

We now show how FCA was applied to our data. The experiments described here are based on the verbs in the data set **vn_restricted** (Section 3.2), i.e. for these experiments we used roughly 2200 French verbs occurring in a restricted number of

[20]Ideally they should have been, since they are all representatives of different classes, but we did not include all diathesis alternations in this simple example.

translated Verbnet classes. We explored classifications built with the feature sets described in Section 3.1.3, that is using subcategorisation frames only (**scf**), and subcategorisation frames combined with additional syntactic (**scf-syn**), semantic (**scf-sem**) or syntactic and semantic (**scf-syn-sem**) features. The rest of this section is structured as follows. After showing how the concept lattice is built, we performed a first series of experiments to determine the best way to filter the concept lattice based on concept selection indices. In the next series of experiments we addressed the association of groups of verbs with thematic role sets. We determined the best performing feature sets and the best way of translating English Verbnet classes.

In our first experiment we only used subcategorisation frames as features. Using the Galicia Lattice Builder software[21], we built a concept lattice based on the formal context $\langle V, F, R \rangle$ such that:

- V is the set of verbs that are either in the translated Verbnet classes (Sec. 3.1.2.1) or in the gold (Sec. 3.2.1.1) and which are also present in our subcategorisation lexicon. We ignore verbs with only one subcategorisation frame as they will result in classes associating verbs with a unique frame.

- F is the set of subcategorisation frames present in the subcategorisation lexicon,

- R is the mapping such that $(v, f) \in R$ iff the subcategorisation lexicon associates the verb v with the SCF f.

The resulting formal context consisted of 2091 objects (verbs) and 238 attributes (frames), giving rise to a lattice of 12802 concepts.

Clearly not all these concepts are interesting verb classes. Classes aim to factorise information and express generalisations about verbs. Hence, concepts with few (1 or 2) verbs can hardly be viewed as classes and similarly, concepts with few frames are less interesting.

To select from this lattice those concepts which are most likely to provide the most relevant verb-frame associations, we used three indices for concept selection: *concept stability*, *separation* and *probability* which have been proposed and analysed in [Klimushkin *et al.*, 2010].

In the next section we explore which of these indices performs best in the context of our application.

[21] `http://www.iro.umontreal.ca/~galicia/`

3.3.1.2 Filtering concept lattices.

[Klimushkin *et al.*, 2010] propose three indices for selecting relevant concepts in concept lattices built on noisy data: *concept stability, separation* and *probability.*

Concept stability is a measure which helps discriminating potentially interesting patterns from irrelevant information in a concept lattice based on possibly noisy data. The stability of a concept $C = (V, F)$ is the proportion of subsets of the extent V which have the same attribute set F as V:

$$\sigma((V,F)) = \frac{|\{A \subseteq V \mid A' = F\}|}{2^{|V|}}. \tag{3.1}$$

Intuitively, a more stable concept is less dependant on any individual object in its extent and is therefore more resistant to outliers or other noisy data items.

Concept separation indicates the significance of the difference between the objects covered by a given concept from other objects and, simultaneously, between its attributes and other attributes:

$$\mathfrak{s}((V,F)) = \frac{|V|\,|F|}{\sum_{v\in V}|\{v\}'| + \sum_{f\in F}|\{f\}'| - |V|\,|F|}. \tag{3.2}$$

Intuitively we expect a concept with high separation index to better sort out the verbs it covers from other verbs and simultaneously the frames it covers from other frames. Whereas concept stability is a measure concerned with either objects or attributes, separation gives information about objects and attributes at the same time.

Concept probability. For an attribute $a \in A$, the attribute set, we denote by p_a the probability of an object to have the attribute a. In practice it is the proportion of objects having a: $p_a = \frac{|\{a\}'|}{|O|}$, where O denotes the set of objects. For $B \subseteq A$, we define p_B as the probability of an arbitrary object having all attributes from B: $p_B = \prod_{a\in B} p_a$. This formulation assumes the mutual independence of attributes. Based on this, and denoting $n = |O|$ we obtain the following formula for the probability of

B being closed:

$$p(B = B'') = \sum_{k=0}^{n} p(|B'| = k, B = B'') \tag{3.3}$$

$$= \sum_{k=0}^{n} \left[\binom{n}{k} p_B^k (1 - p_B)^{n-k} \prod_{a \notin B} (1 - p_a^k) \right] \tag{3.4}$$

A small $p(B = B'')$ suggests a small probability of the attribute combination B to be a concept intent by chance only (and $p(B = B'') \approx 1$ that there is a high probability that the combination is a concept intent by chance). However, this reasoning is based on the independence of the attributes, which in our particular case can not be warranted.

Computing stability, separation and probability indices.

Stability. Calculating stability is known to be #P-complete ([Kuznetsov, 2007]), however [Roth *et al.*, 2006] show that when the concept lattice is known it can be computed efficiently by a bottom-up traversal algorithm. This is the algorithm we used to compute concept stability.

Separation can be computed in $\mathcal{O}(|O| + |A|)$ time, where O and A are the object and attribute sets respectively. Computing separation is the least prohibitive of the three indices.

Probability. [Klimushkin *et al.*, 2010] show that computing probability of only one concept involves $\mathcal{O}(|O|^2 \cdot |A|)$ multiplication operations which is computationally very costly. We therefore computed approximations derived as follows. First, we consider $\prod_{a \in B} (1 - p_a^k) \approx 1$ for $k > 40$. In view of this, Equation (3.4) becomes:

$$p(B = B'') = \sum_{k=0}^{40} \left[\binom{n}{k} p_B^k (1 - p_B)^{n-k} \prod_{a \notin B} (1 - p_a^k) \right] \tag{3.5}$$

$$+ \sum_{k=41}^{n} \left[\binom{n}{k} p_B^k (1 - p_B)^{n-k} \right] \tag{3.6}$$

As $\sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} = 1$, Term (3.6) can be rewritten as:

$$1 - \sum_{k=0}^{40} \left[\binom{n}{k} p_B^k (1-p_B)^{n-k} \right] = \tag{3.7}$$

$$1 - F(40; n, p_B). \tag{3.8}$$

where $F(k; n, p) = \sum_{i=0}^{k} \binom{n}{i} p^i (1-p)^{n-1}$ is the cumulative distribution function of the binomial distribution[22] and can be computed using various statistical software packages. Term (3.5) can also be computed more easily considering that $\binom{n}{k} p_B^k (1-p_B)^{n-k}$ are binomial densities the computation of which is also provided by statistics software[23].

In the following we measure the performance of the three concept selection indices with respect to our data. The experimental setting is as follows. We first selected a number of N (1500) concepts with best selection index. The selected concepts were aligned with the classes translated from Verbnet (with the svm method, see Section 3.1.2.1): For each translated class, we selected the concept with best precision/recall f-measure. For a translated Verbnet class C_{VN} and an FCA concept C_{FCA} we computed recall R, precision P and f-measure F as follows:

$$R = \frac{|C_{\text{VN}} \cap C_{\text{FCA}}|}{|C_{\text{VN}}|}, P = \frac{|C_{\text{VN}} \cap C_{\text{FCA}}|}{|C_{\text{FCA}}|}, F = \frac{2RP}{R+P}$$

We then associated to the concept with best f-measure the theta grid of the translated Verbnet class. Next we compared the obtained verb, theta grid associations with those given by a reference. As for our task recall was more important than precision, we used the $F2$ measure, which gives more weight to recall, for comparison.

As reference we used the data used for training the classifier for learning the translated Verbnet classes (see Section 3.1.2.1). We chose this as reference rather than the gold standard (cf. Section 3.2.1.1) because here we were evaluating the concept selection indices: we wanted to check which index selected the most relevant concepts, that is those best matching the translated classes. The reference consisted of the ⟨verb, theta role⟩ pairs marked as positive examples in the training set.

Table 3.12 shows the F2 scores and coverage when using only one index at a time. For stability and separation we applied the method above on the top ranking

[22] Source Wikipedia: `http://en.wikipedia.org/wiki/Binomial_distribution`

[23] We used the R software environment for statistical computing (`http://www.r-project.org/`)

	cov.	prec.	rec.	F	F2
stab only	39.88	18.96	32.55	24.00	26.27
sep only	34.25	28.37	21.52	24.47	23.41
prob only	35.53	26.60	20.73	23.30	22.38
w/o filtering	100	12.30	60.96	20.47	26.30

Table 3.12: F2 scores and coverage for stability, separation and the 6th probability 10-quantile.

1500 concepts. Regarding probability, at first sight, we should consider best the concepts with lowest probability – because the probability of their intents of being closed by chance only is accordingly low. However, looking at the data we found that these concepts have very few verbs and large intent (frame) sets - which rather suggest improbable or rare verb groups. On the other hand, the interpretation of concept probability suggests that a concept with a probability close to 1 could occur by chance only. For these reasons, to assess probability separately we settled on the 6th 10 quantile. The results confirmed the observations of [Klimushkin *et al.*, 2010]: stability alone gave F2 scores close to an upper bound – the results obtained without filtering, i.e. aligning the translated classes with all the concepts of the lattice. The results for separation and probability were several points lower.

As we only selected $\sim$ 10% of the total number of concepts we also had to make sure that the selected concepts covered at least a reasonable amount of verbs. Table 3.12 shows that for all indices the coverage is unsatisfactory.

Combining stability, separation and probability. [Klimushkin *et al.*, 2010] investigate the performance of the stability, separation and probability indices at finding the original concepts in lattices produced from contexts which were previously altered by introducing two types of noise:

Type I noise is obtained by altering every cell in the context with some probability,

Type II noise is obtained by adding a given number or proportion of random objects or attributes.

According to this, our contexts are affected by Type I noise rather than Type II. [Klimushkin *et al.*, 2010] found that stability was most effective at sorting out Type II noise, but also proved helpful in the case of Type I noise. In contrast, they suggest that separation and probability can not be used on their own but should rather serve as a normalising measure for stability. The most promising combination seemed to

be

$$\text{stability} + k_1 \cdot \text{separation} - k_2 \cdot \text{probability}, \text{ for some constants } k_1 \text{ and } k_2.$$

In the following we started from the assumption that the most effective index for selecting relevant concepts is given by a linear combination of stability, separation and probability:

$$k_1 \cdot \text{stability} + k_2 \cdot \text{separation} - k_3 \cdot \text{probability}$$

and empirically determined the coefficients k_1, k_2 and k_3 such that the selected concepts perform best with respect to our task.

We proceeded as follows. For a fixed k_1, k_2 and k_3 combination we computed the corresponding linear combination for the concepts and select the 1500 concepts ranking highest. As in the previous experiments, we measured the performance of the selected concepts by aligning them with the translated Verbnet classes and by comparing the alignments with the same reference as before. We considered the "best" k_1, k_2, k_3 combination the one giving highest F2 scores and good coverage.

The upper part of Table 3.13 shows the results for a first series of experiments where k_1 and k_2 were assigned the values 0.5 and 1 and k_3 0.25 and 0.5 (The lines are sorted by decreasing F2 score). They suggest that the stability and separation coefficients had less impact on coverage and F2 score than the probability coefficient. Interestingly the coverage was correlated with the F2 score.

In the second series of experiments, shown in the lower part of Table 3.13, we kept the stability and separation coefficients fixed and varied only the probability coefficient. The results suggested that the probability coefficient may not help at selecting the most relevant concepts in our setting. This may be due first to the fact that our attributes are not independent (we assumed independence of attributes when setting up the formula for computing the probability index) and second to the fact that we had to approximate the probability index and this approximation may not be accurate enough.

In the next series of experiments, shown in Table 3.14, we investigated the impact of the number of preselected concepts (500). The results show that with this smaller number of concepts the selected concepts reached a slightly smaller F2 score but a substantially lower coverage. Also, in this configuration the probability index did seem to be helpful. The results for the next set of data, where we preselected 1000 concepts are shown in Table 3.15. The F2 score and coverage were only slightly lower than when preselecting 1500 concepts and again the probability index seemed

k_1, k_2, k_3	**cov.**	**prec.**	**rec.**	**F**	**F2**
1, 1, 0.25	98.04	11.87	55.19	19.53	24.89
1, 0.5, 0.25	98.04	11.87	55.19	19.53	24.89
1, 0.5, 0.5	57.69	17.08	30.18	21.82	24.04
1, 1, 0.5	56.15	17.45	29.13	21.82	23.82
0.5, 0.5, 0.25	56.15	17.45	29.13	21.82	23.82
0.5, 1, 0.25	53.81	18.03	27.82	21.88	23.36
0.5, 0.5, 0.5	49.72	18.55	26.25	21.73	23.06
0.5, 1, 0.5	49.90	18.61	25.98	21.67	22.95
1, 1, 0	98.04	12.05	55.12	19.78	25.16
1, 1, 0.05	98.04	12.05	55.12	19.78	25.16
1, 1, 0.005	98.04	12.05	55.12	19.78	25.16
1, 1, 0.0005	98.04	12.05	55.12	19.78	25.16
1, 1, 0.1	98.00	11.91	55.38	19.60	25.00
1, 1, 0.2	98.08	11.88	55.12	19.53	24.91
1, 1, 0.25	98.04	11.87	55.12	19.53	24.89
1, 1, 0.3	98.00	11.79	55.38	19.44	24.80
1, 1, 0.4	59.95	16.27	31.23	21.40	23.91
1, 1, 0.5	56.16	17.45	29.13	21.83	23.82

Table 3.13: F2 scores and coverage for various k_1, k_2, k_3 combinations.

k_1, k_2, k_3	**cov.**	**prec.**	**rec.**	**F**	**F2**
1, 1, 0.25	46.23	18.87	27.29	22.32	23.76
1, 1, 0.3	45.46	18.51	26.77	21.89	23.30
1, 1, 0.2	47.72	18.25	26.77	21.33	23.16
1, 0.5, 0.25	47.21	17.74	26.77	21.34	22.89
1, 1, 0	97.61	10.50	55.64	17.67	22.87
0.5, 0.5, 0.5	39.58	20.67	24.15	22.28	22.87
0.5, 1, 0.5	39.54	20.54	24.15	22.20	22.81
1, 0.5, 0.5	42.05	20.82	23.88	22.25	22.77
0.5, 1, 0.25	42.27	21.23	23.62	22.36	22.77
1, 1, 0.5	42.01	21.08	23.62	22.28	22.71
0.5, 0.5, 0.25	42.01	21.08	23.62	22.28	22.71
1, 1, 0.4	43.84	19.62	24.41	21.75	22.57
1, 1, 0.1	73.03	10.67	44.09	17.19	21.58
1, 0, 0	25.39	26.73	19.16	22.32	21.16
0, 1, 0	22.24	24.39	7.87	0.12	10.17

Table 3.14: F2 scores and coverage for various k_1, k_2, k_3 combinations when selecting from the top ranked 500 concepts.

to have only low relevance for the overall results.

In the first place, these experiments suggested that the best linear combination is the sum of the stability and separation indices. Second, it does not seem evident that probability had a positive effect on the relevance of the selected concepts. However, it did improve F-measure when the number of selected concepts is lower (500 or 1000 vs. 1500 in our experiments). Hence it is probably a better strategy to select a larger

k_1, k_2, k_3	cov.	prec.	rec.	F	F2
1, 1, 0.1	97.74	11.42	55.38	18.93	24.25
1, 1, 0	97.86	11.33	55.12	18.79	24.08
1, 1, 0.3	59.14	16.06	31.50	21.27	23.86
0, 1, 0	31.96	26.98	22.31	24.43	23.68
1, 1, 0.2	97.19	10.90	56.69	18.29	23.62
0.5, 0.5, 0.5	47.89	19.39	26.51	22.39	23.62
0.5, 1, 0.25	49.34	19.22	25.98	22.09	23.26
1, 1, 0.25	83.98	11.12	50.39	18.22	23.15
1, 0.5, 0.5	50.79	17.99	26.25	21.34	22.76
1, 1, 0.4	51.43	17.89	26.25	21.28	22.71
0.5, 1, 0.5	45.08	21.95	23.10	22.51	22.70
0.5, 0.5, 0.25	50.62	17.79	26.25	21.21	22.66
1, 1, 0.5	50.62	50.62	26.25	21.21	22.66
1, 0.5, 0.25	90.24	10.61	46.19	17.26	21.82
1, 0, 0	29.57	26.74	19.16	22.32	21.16

Table 3.15: F2 scores and coverage for various k_1, k_2, k_3 combinations when selecting from the top ranked 1000 concepts.

number of concepts (1500) and not take probability into account, more so as we first had to use an approximation to compute it, which may be too rough and second the computation of probability is based on the independence of attributes which is not warranted in our case.

In the following experiments we adopted this strategy and selected the 1500 concepts where the sum of the stability and separation indices was highest.

3.3.1.3 Associating FCA concepts with thematic role sets.

To associate FCA concepts with thematic role sets we aligned them with the translated Verbnet classes produced as shown in Section 3.1.2.1. For each translated Verbnet class we identified among the 1500 filtered FCA concepts the one(s) with best f-measure between precision and recall. For a translated Verbnet class C_{VN} (consisting of French verbs) and the extent (verb set) of an FCA concept C_{FCA} *precision*, *recall* and *F-measure* were computed as follows:

$$R = \frac{|C_{\text{VN}} \cap C_{\text{FCA}}|}{|C_{\text{VN}}|}, P = \frac{|C_{\text{VN}} \cap C_{\text{FCA}}|}{|C_{\text{FCA}}|}, F = \frac{2RP}{R+P} \tag{3.9}$$

The translated Verbnet class was then associated with this FCA concept(s). Thus the verbs in the FCA concept were effectively associated with the thematic grid of the translated class and in the same time with the syntactic frames in the intent (attribute set) of the FCA concept. As shown in Section 3.1.2.1, the translated Verbnet classes were obtained using different methods.

The aim of the next series of experiments was two fold: First to assess which of the translation methods presented in Section 3.1.2.1 is most appropriate and second, to explore the performance of the feature sets shown in Section 3.1.3. To investigate these questions, we used FCA to compute classifications grouping verbs and frames based on different feature sets combining subcategorisation frames and additional syntactic and/or semantic features. We associated the groups of verbs with thematic role sets by aligning them with translated Verbnet classes obtained using the different translation methods described in Section 3.1.2.1. Finally, we compared the ⟨verb, thematic grid⟩ associations induced by the classification with the gold standard by [Sun *et al.*, 2010] (see Section 3.2.1.2) in terms of precision, recall and f-measure.

To sum up, the feature sets used in these experiments are the following (see Section 3.1.3 for a more detailed description): **scf** (subcategorisation frames only), **scf-synt**, (subcategorisation frames and syntactic features), **scf-sem**, (subcategorisation frames and semantic features) and **scf-synt-sem** (subcategorisation frames with both syntactic and semantic features). The translated Verbnet classes were obtained with the following methods (see Section 3.1.2.1 for details): **median** (classes were translated based on translation frequencies), **svm** (classes were translated with SVM, we selected 6000 ⟨verb, class⟩ pairs with highest probability estimates), **svm-best** (classes were translated with SVM, we selected for each verb, the ⟨verb, class⟩ pair with highest probability estimate) and **svm-median** (classes were translated with SVM, we selected for each verb, those ⟨verb, class⟩ pairs with probability estimates above median).

Table 3.16 shows the results for the various experiments, ordered by decreasing f-measure. These results suggest that the best performing configuration for our application is to translate the Verbnet classes using the **svm** method and to build the FCA concepts based on the **scf-sem** attribute set (that is to use subcategorisation frames and additional semantic features (Table 3.5b) as attributes). Figure 3.7 shows the associations between concepts, thematic grids and frames generated by our method using this setting. The figure shows the concepts which were associated to a thematic grid in the gold standard, and for each of these concepts, their attribute set (syntactic frames and additional features), the associated thematic grid(s), the number of verbs in the concept and the hierarchical relations between the concepts as given by the concept lattice. The method selected 10 concepts which were labeled with the 11 Verbnet classes (theta grids) occuring in the gold. 9 FCA concepts were mapped to exactly one Verbnet class and 1 was associated with 2 Verbnet classes.

We showed in [Falk *et al.*, 2010] and [Falk and Gardent, 2010], where we used FCA in the same way to bootstrap a classification of French verbs, that the resulting

Method	cov.	prec	rec	f
svm, scf-sem	96.17	24.09	75.00	36.47
svm, scf-sem	96.05	23.95	75.00	36.31
svm-best, scf-sem	91.59	23.86	75.00	36.21
svm-best, scf-synt-sem	92.47	23.56	73.21	35.65
svm, scf	95.37	23.48	73.80	35.63
svm, scf-synt	96.34	21.51	74.40	33.38
svm-best, scf-synt	92.60	21.48	74.40	33.33
median, scf-sem	92.52	20.36	80.36	32.49
median, scf-synt-sem	92.38	20.40	79.76	32.48
median, scf	93.26	18.72	83.33	30.57
svm-best, scf	95.18	20.78	57.14	30.48
median, scf-synt	93.29	18.49	81.55	30.14
svm-median, scf-synt-sem.	91.86	16.85	82.14	27.96
svm-median, scf-sem	89.26	15.03	73.21	24.95
svm-median, scf-synt	92.13	14.12	70.83	23.54
svm-median, scf	95.62	12.07	50.00	19.44

Table 3.16: Verb coverage and precision, recall and f-measure for produced ⟨verb, theta⟩ associations wrt. the gold standard. The translated classes are produced either using a machine learning method (*svm*) or by simply collecting the most frequent translations (*median*). When building the translated classes with svm, we select either the 6000 best scoring verb, class pairs (*svm*) or for each verb the best scoring ⟨verb, class pair⟩ (*svm-best*) or for each verb the ⟨verb, class⟩ pairs with scores above the median (*svm-median*). At the construction of the lattice we use frames only (*scf*), additional syntactic features (*scf-syn*), additional semantic features (*scf-sem*) or both (*scf-synt-sem*). We pre-selected 1500 concepts with best sum of stability and separation indeces.

classifications have good factorisation power. A partial qualitative evaluation suggested that the classes built this way adequately describe the association between verb sets, syntactic frames and thematic grids.

In Chapters 4 and 5 we will present a detailed evaluation of this classification method (i) with respect to the created groups of verbs and their associations to the thematic role sets and (ii) with respect to the associations of the groups of verbs with both syntactic frames and thematic role sets. In these evaluations we will use FCA classifications built with the configuration which proved to perform best in the experiments presented here, which we recall in the following.

> We found that the best performing configuration for building Verbnet like classes of French verbs with FCA is the following. The attribute (feature) set consists of subcategorisation frames and additional semantic features (**scf-sem** feature set). We used the sum of the stability and separation concept selection indices to select 1500 concepts. From these we kept those where the proportion of verbs shared with a translated Verbnet class is best. The Verbnet classes were translated using the **svm** method.

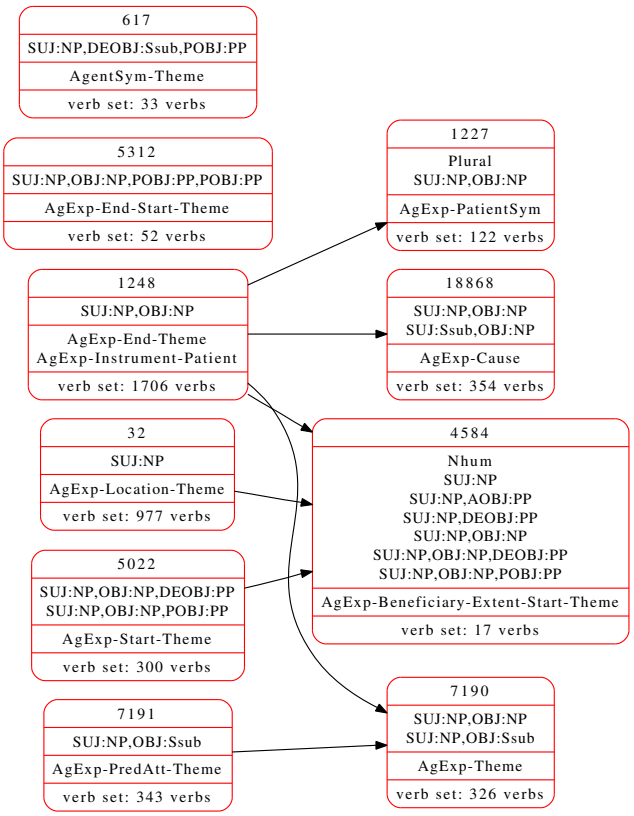


Figure 3.7: French verb ↔ synt. frames ↔ theta grid associations obtained with FCA based on **scf-sem** features and Verbnet classes translated using the **svm** method.

3.3.2 Incremental Growing Neural Gas with Feature Maximisation (IGNGF)

The second clustering technique we used in the following experiments is a neural clustering algorithm called Incremental Growing Neural Gas with Feature Maximisation (IGNGF). We start with a general introduction of this method and then describe how it is applied to acquire French verb classes[24].

Like other neural free topology methods such as Neural Gas (NG) [Martinetz and Schulten, 1991], Growing Neural Gas (GNG) [Fritzke, 1995], or Incremental Growing Neural Gas (IGNG) [Prudent and Ennaji, 2005], the IGNGF method makes use of Hebbian learning [Hebb, 1949] for dynamically structuring the learning space. Hebbian learning is inspired by a theory from neurosciences which explains how neurons connect to build neural networks. In the following we briefly discuss the characteristics of the clustering techniques listed above. More detailed explanations

[24]Parts of this section and of Section 4.2.2 are published as [Falk *et al.*, 2012]

can be found in [Lamirel *et al.*, 2011b].

Whereas for NG the number of output clusters is fixed, GNG adapts the number of clusters during the learning phase, guided by the characteristics of the data to be classified. Clusters and connections between them can be created or removed depending on evolving characteristics of learning (as for example the "age" or "maturity" of connections and the cumulated error rate of each cluster prototype). A drawback of this approach is that clusters are created or removed after a fixed number of iterations yielding clusters which might not appropriately represent complex or sparse multidimensional data. With the IGNG clustering method this issue is addressed by allowing more flexibility when creating new clusters: a cluster is added whenever the distance of a new data point to an existing cluster is above a predefined global threshold, the average distance of all the data points to the centre of the data set. The clustering process thus becomes incremental: each incoming data point (verb in our setting) is considered as a potential cluster. At each iteration over all the data points, a data point is connected with the "closest" clusters and at the same time interacts with the existing clustering by strengthening the connections between these "closest" clusters and weakening those to other, less related clusters. Because of these dynamically changing interactions between clusters, these methods are "winner take most" methods in contrast to K-means, which represents a "winner-take-all" method. The notion of "closeness" is based on a distance function computed from the features associated to the data points.

3.3.2.1 Clustering based on feature maximisation.

IGNGF uses the Hebbian learning process as IGNG, but the use of a standard distance measure as adopted in IGNG for determining the "closest" cluster is replaced in IGNGF by feature maximisation. Feature maximisation is a cluster quality metric which favours clusters with maximum feature F-measure. *Feature F-measure* (FF) is the harmonic mean of *feature recall* (FR) and *feature precision* (FP) which in turn are defined as[25]:

$$FR_c(f) = \frac{\sum_{v \in c} W_v^f}{\sum_{c' \in C} \sum_{v \in c'} W_v^f}, \qquad FP_c(f) = \frac{\sum_{v \in c} W_v^f}{\sum_{f' \in F_c, v \in c} W_v^{f'}}$$

[25]Since *feature recall* is equivalent to the conditional probability $P(c|p)$ and *feature precision* is equivalent to the conditional probability $P(p|c)$, this former strategy can be classified as an expectation maximisation approach with respect to the original definition given by Dempster and al. [Dempster *et al.*, 1977].

where W_x^f represents the weight of the feature f for element x and F_c designates the set of features associated with the verbs occuring in cluster c. A feature is then said to be maximal for a given cluster iff its feature F-measure is higher for that cluster than for any other cluster. Finally the Feature F-measure FF_c of a cluster $c \in C$ is the average of the Feature F-measures of the maximal features for c:

$$FF_c = \frac{\sum_{f \in F_c} FF_c(f)}{|F_c|}$$

With feature maximisation, the clustering process is roughly the following. An incoming data point v is temporary added to every existing cluster, its feature profile is updated (i.e. each cluster is associated with its maximal features) and its average Feature F-measure is computed. Then the winning cluster is the cluster which maximises the distance κ given in Equation (3.10).

$$\kappa(c) = \Delta(FF_c) * |F_c \cap F_v| - \frac{EucDist(\vec{c}, v)}{weight} \tag{3.10}$$

where $\Delta(FF_c)$ represents the gain in Feature F-measure for the new cluster and $F_c \cap F_v$ are the features shared by cluster c and the data point v. This way, those clusters are preferred which share more features with the new data point and clusters which don't have any common feature with the data point are ignored. The gain in Feature F-measure multiplied by the number of shared features is adjusted by the euclidean distance of the new data point v to the cluster centroid vector $\vec{c}$. Thus, the smaller the euclidean distance to the cluster, less the κ value decreases. The influence of the euclidean distance can be parametrised with a *weight* factor. Clusters with negative κ score are ignored. The data point is then added to the cluster c with maximal $\kappa(c)$ and the connections to the winner and its neighbours are updated.

The IGNGF method was shown to outperform other usual neural and non neural methods for clustering tasks on relatively clean data [Lamirel *et al.*, 2011b]. Since we use features which are extracted from manually validated sources and are therefore less noisy than features extracted from corpora, IGNGF is a good candidate for our task.

To build our clusterings, we applied an IDF-Norm weighting scheme [Robertson and Sparck Jones, 1976]. In this scheme the IDF component (3.11) decreases the influence of the most frequent features.

$$idf(f) = log(\frac{\#verbs}{freq(f)}) \tag{3.11}$$

Normalisation is used to compensate for discrepancies in feature number. It consists in dividing the elements of each feature vector of a verb by the number of features the verb has.

3.3.2.2 Cluster labeling.

The primary goal of the IGNGF method is to cluster the data points (verbs). Cluster labeling is the subsequent process of associating these clusters with the features, which are considered most characteristic for the clusters with respect to (i) the clustering and (ii) the application at hand.

The feature maximisation and cluster labeling performed by the IGNGF method has proven promising both for visualising or synthesising clustering results [Lamirel *et al.*, 2008] and for validating or optimising a clustering method [Attik *et al.*, 2006]. We make use of these processes in all our experiments (with IGNGF of course but also with K-Means) and systematically compute cluster labeling on the output clusterings. Cluster labeling facilitates clustering interpretation in that it clearly indicates the association between clusters (verbs) and their prevalent features. And it supports the creation of a Verbnet style classification in that it directly provides classes associating groups of verbs with thematic grids and subcategorisation frames.

The most prominent candidate features for labeling a cluster clearly are the cluster maximising features, because these are the features which had the greatest impact on the creation of the clustering in the first place. With this labeling scheme, a feature may be associated with at most one cluster. But, as we shall see in Chapter 5, this cluster labeling did not yield adequate ⟨verb, subcategorisation frame⟩ associations. Therefore we also experimented with a second cluster labeling method which consists in associating the clusters (in addition to the cluster maximising features) with the features having a Feature F-measure above a global threshold. In our application this global threshold is the average of the Feature F-measure of the cluster maximising features. These features were less relevant to the clustering algorithm, but were important for the interpretation of the clusters as syntactic-semantic verb classes. In the following we will refer to the first cluster labeling method, where clusters are labeled with the cluster maximising features as **Fmax** and to the second as **Fpos**.

3.3.2.3 Clustering French verbs.

In this section, we show how we apply the IGNGF clustering method to build a syntactic-semantic classification of French verbs and determine the configuration

yielding the best performing IGNGF clustering. We also show a sample resulting cluster and perform a short qualitative discussion. All clusterings using IGNGF were computed by Jean-Charles Lamirel.

The objects to be clustered were the French verbs in the data set **vn_restricted**, that is we clustered roughly 2200 French verbs occuring in a restricted number of translated Verbnet classes and in our merged syntactic lexicon, described in Section 3.1. The translated classes were obtained with the **svm** method (see Section 3.1.2.1. We used these translated classes because they were shown to perform best in our experiments with the FCA classification method, discussed in Section 3.3.1.3). We built clusterings based on the feature sets described in Section 3.1.3. In the simplest case the features were the subcategorisation frames (**scf**) associated to the verbs by our lexicon. We also experimented with feature sets combining subcategorisation frames with (i) additional syntactic and semantic features extracted from the lexicons and (ii) features derived from the translated Verbnet classes (**grid**). To sum up, the feature sets we used in building the clusterings are the following: **scf** (subcategorisation frames only), **scf-synt** (subcategorisation frames and additional syntactic features), **scf-sem** (subcategorisation frames and additional semantic features), **scf-synt-sem** (subcategorisation frames and both additional syntactic and semantic features), **grid-scf** (subcategorisation frames and thematic grids), **grid-scf-synt** (subcategorisation frames, additional syntactic features and thematic grids), **grid-scf-sem** (subcategorisation frames, additional semantic features and thematic grids) and **grid-scf-synt-sem** (subcategorisation frames, both additional syntactic and semantic features and thematic grids). The feature sets were in essence the same as those used for building the FCA concepts. The only difference is that in the FCA experiments we did not use the features derived from translated Verbnet classes. To determine which of these feature sets is most helpful in this clustering task, we used each of them to perform the clustering, and compared the results with the V-gold standard (proposed by [Sun *et al.*, 2010] and presented in Section 3.2.1.1). For this comparison we used the mPUR and ACC clustering evaluation measures which were also used by [Sun *et al.*, 2010]. We will give a more detailed explanation of these measures in Chapter 4, where we also go more in detail with respect to this evaluation.

To determine the number of clusters, we let this number vary between 1 and 30 to obtain a partition that reached an optimum F-measure and a number of clusters that was in the same order of magnitude as the initial number of gold classes (i.e. 11 classes). Table 3.17 shows the results of this evaluation. The table shows that the best results were obtained using the **grid-scf-sem** feature set (with an F-measure

Feat. set	Nbr. feat.	Nbr. verbs	mPUR	ACC	F (Gold)	Nbr. classes
scf	220	2085	0.93	0.48	0.64	17
grid-scf	231	2085	0.94	0.54	0.68	13
grid-scf-sem	237	2183	0.86	0.59	**0.70**	13
grid-scf-synt	236	2150	0.87	0.50	0.63	14
grid-scf-synt-sem	242	2201	0.99	0.52	0.69	16
scf-sem	226	2183	0.83	0.55	0.66	23
scf-synt	225	2150	0.91	0.45	0.61	15
scf-synt-sem	231	2101	0.89	0.47	0.61	16

Table 3.17: IGNGF clustering results for various feature sets, consisting of a combination of subcategorisation frames *scf*, syntactic (*synt*) and semantic (*sem*) features and thematic grids (*grid*).

Method	mPUR	ACC	F (Gold)	Nbr. classes	Cov.	CMP
IGNGF with IDF and norm.	0.86	0.59	0.70	13	0.72	0.30
no IDF	0.86	0.59	0.70	17	0.81	0.18
no norm.	0.78	0.62	0.70	18	0.72	0.15
no IDF, no norm.	0.87	0.55	0.68	14	0.81	0.21

Table 3.18: The impact of the IDF-norm weighting scheme.

of 0.70). In the following we therefore used IGNGF with this feature set. Table 3.18 shows the impact of IDF feature weighting and feature vector normalisation on the clustering. The benefit of preprocessing the data appears clearly. When neither IDF weighting nor vector normalisation are used, F-measure decreased from 0.70 to 0.68 and cumulative micro-precision from 0.30 to 0.21. When either normalisation or IDF weighting is left out, the cumulative micro-precision dropped by up to 15 points (from 0.30 to 0.15 and 0.18). Cumulative micro-precision is an unsupervised clustering metric which evaluates the quality of a clustering with regard to the cluster features rather than with respect to a gold standard. It will be introduced and discussed in Chapter 4. By consequence we used both IDF weighting and feature vector normalisation in the following clustering experiments. Table 3.19 shows a sample output cluster produced by the IGNGF clustering method using the **grid-scf-sem** feature set and the IDF-norm weighting scheme. The clusters were labeled according to the **Fmax** labeling scheme. Maximised features are displayed in decreasing order and features whose Feature F-measure is under the average Feature F-measure of the overall clustering are clearly delineated from others (by two rows of stars). In addition, for each verb in a cluster, a confidence score is displayed which is the ratio between the sum of the F-measures of its cluster maximised features over the sum of the F-measures of the overall cluster maximised features. Verbs whose confidence score is 0 are considered as orphan data.

The following preliminary qualitative evaluation shows that the IGNGF clus-

C6- 14(14) [197(197)]
———-
Prevalent Label — = AgExp-Cause

0.341100 G-AgExp-Cause
0.274864 C-SUJ:Ssub,OBJ:NP
0.061313 C-SUJ:Ssub
0.042544 C-SUJ:NP,DEOBJ:Ssub
****
****
0.017787 C-SUJ:NP,DEOBJ:VPinf
0.008108 C-SUJ:VPinf,AOBJ:PP
...
[**déprimer 0.934345 4(0)] [affliger 0.879122 3(0)] [éblouir 0.879122 3(0)] [choquer 0.879122 3(0)] [décevoir 0.879122 3(0)] [décontenancer 0.879122 3(0)] [décontracter 0.879122 3(0)] [désillusionner 0.879122 3(0)] [**ennuyer 0.879122 3(0)] [fasciner 0.879122 3(0)] [**heurter 0.879122 3(0)] ...

Table 3.19: Sample output for a cluster produced with the **grid-scf-sem** feature set and the IGNGF clustering method.

tering combined with its cluster labeling techniques produced indeed a coherent syntactic-semantic verb classification. We carried out a manual analysis of the clusters examining both the semantic coherence of each cluster (do the verbs in that cluster share a semantic component?) and the association between the thematic grids, the verbs and the syntactic frames provided by clustering.

Semantic homogeneity. To assess semantic homogeneity, we examined each cluster and seeked to identify one or more Verbnet labels characterising the verbs contained in that cluster. From the 13 clusters produced by clustering, 11 clusters could be labeled. Table 3.20 shows these eleven clusters, the associated labels (abbreviated Verbnet class names), some example verbs, a sample subcategorisation frame drawn from the cluster maximising features and an illustrating sentence. As can be seen, some clusters group together several subclasses and conversely, some Verbnet classes are spread over several clusters. This is not necessarily incorrect though. To start with, recall that we are aiming for a classification which groups together verbs with the same thematic grid. Given this, cluster C2 correctly groups together two Verbnet classes (other_cos-45.4 and hit-18.1) which share the same thematic grid (cf. Table 3.9). In addition, the features associated with this cluster indicate that verbs in these two classes are transitive, select a concrete object, and can be pronominalised which again is correct for most verbs in that cluster. Similarly, cluster C11 groups together verbs from two Verbnet classes with identical theta grid (light_emission-43.1 and modes_of_being_with_motion-47.3) while its associated features correctly indicate that verbs from both classes accept both the intransitive form without object

(*la jeune fille rayonne / the young girl glows, un cheval galope / a horse gallops*) and with a prepositional object (*la jeune fille rayonne de bonheur / the young girl glows with happiness, un cheval galope vers l'infini / a horse gallops to infinity*). The third cluster grouping together verbs from two Verbnet classes is C7 which contains mainly judgement verbs (to applaud, bless, compliment, punish) but also some verbs from the (very large) other_cos-45.4 class. In this case, a prevalent shared feature is that both types of verbs accept a de-object that is, a prepositional object introduced by "de" (*Jean applaudit Marie d'avoir dansé / Jean applaudit Marie for having danced; Jean dégage le sable de la route / Jean clears the sand of the road*). The semantic features necessary to provide a finer grained analysis of their differences are lacking.

Interestingly, clustering also highlights classes which are semantically homogeneous but syntactically distinct. While clusters C6 and C10 both contain mostly verbs from the amuse-31.1 class (*amuser, agacer, énerver, déprimer*), their features indicate that verbs in C10 accept the pronominal form (e.g., *Jean s'amuse*) while verbs in C6 do not (e.g., **Jean se déprime*). In this case, clustering highlights a syntactic distinction which is present in French but not in English. In contrast, the dispersion of verbs from the other_cos-45.4 class over clusters C2 and C7 has no obvious explanation. One reason might be that this class is rather large (361 verbs) and thus might contain French verbs that do not necessarily share properties with the original Verbnet class.

Syntax and Semantics. We examined whether the prevalent syntactic features labeling each cluster were compatible with the verbs and with the semantic class(es) manually assigned to the clusters. Table 3.20 sketches the relation between cluster, syntactic frames and Verbnet like classes. It shows for instance that the prevalent frame of the C0 class (manner_speaking-37.3) correctly indicates that verbs in that cluster subcategorise for a sentential argument and an AOBJ (prepositional object in "à") (e.g., *Jean bafouille à Marie qu'il est amoureux / Jean stammers to Mary that he is in love*); and that verbs in the C9 class (characterize-29.2) subcategorise for an object NP and an attribute (*Jean nomme Marie présidente / Jean appoints Marie president*). In general, we found that the prevalent frames associated with each cluster adequately characterise the syntax of that verb class.

As for FCA, we performed a detailed evaluation of the groups of verbs obtained with the IGNGF clustering technique and their associations to sets of subcategorisation frames and thematic roles with respect to both the groups of verbs and their association with thematic roles and syntactic frames. This evaluation is presented in Chapters 4 and 5.

C0	speaking: babiller, bafouiller, balbutier SUJ:NP,OBJ:Ssub,AOBJ:PP *Jean bafouille à Marie qu'il l'aime / Jean stammers to Mary that he is in love*
C1	put: entasser, répandre, essaimer SUJ:NP,POBJ:PP,DUMMY:REFL Loc, Plural *Les déchets s'entassent dans la cour / Waste piles in the yard*
C2	hit: broyer, démolir, fouetter SUJ:NP,OBJ:NP Nhum *Ces pierres broient les graines / These stones grind the seeds.* other_cos: agrandir, alléger, amincir SUJ:NP,DUMMY:REFL *les aéroports s'agrandissent sans arrêt / airports grow constantly*
C4	dedicate: s'engager à, s'obliger à, SUJ:NP,AOBJ:VPinf,DUMMY:REFL *Cette promesse t'engage à nous suivre / This promise commits you to following us*
C5	conjecture: penser, attester, agréer SUJ:NP,OBJ:Ssub *Le médecin atteste que l'employé n'est pas en état de travailler / The physician certifies that the employee is not able to work*
C6	amuse: déprimer, décontenancer, décevoir SUJ:Ssub,OBJ:NP SUJ:NP,DEOBJ:Ssub *Travailler déprime Marie / Working depresses Marie* *Marie déprime de ce que Jean parte / Marie depresses because of Jean's leaving*
C7	other_cos: dégager, vider, drainer, sevrer judgement SUJ:NP,OBJ:NP,DEOBJ:PP *vider le récipient de son contenu / empty the container of its contents* applaudir, bénir, blâmer, SUJ:NP,OBJ:NP,DEOBJ:Ssub *Jean blame Marie d'avoir couru / Jean blames Mary for runnig*
C9	characterise: promouvoir, adouber, nommer SUJ:NP,OBJ:NP,ATB:XP *Jean nomme Marie présidente / Jean appoints Marie president*
C10	amuse: agacer, amuser, enorgueillir SUJ:NP,DEOBJ:XP,DUMMY:REFL *Jean s'enorgueillit d'être roi/ Jean is proud to be king*
C11	light: rayonner,clignoter,cliqueter SUJ:NP,POBJ:PP *Jean clignote des yeux / Jean twinkles his eyes* motion: aller, passer, fuir, glisser SUJ:NP,POBJ:PP *glisser sur le trottoir verglacé / slip on the icy sidewalk*
C12	transfer_msg: enseigner, permettre, interdire SUJ:NP,OBJ:NP,AOBJ:PP *Jean enseigne l'anglais à Marie / Jean teaches Marie English.*

Table 3.20: Relations between clusters, syntactic frames and Verbnet like classes.

3.3.2.4 Associating verb clusters with syntactic frames and thematic grids.

To obtain a Verbnet like classification, the verb clusters produced by IGNGF need to be associated with subcategorisation frames and with thematic role sets. As we saw in Section 3.3.2.2, this clustering method associates each cluster in its output clustering with a ranked list of features that best characterise that cluster. Since the syntactic frames and thematic grids are among the features used, they can be used to "label" the clusters and thus provide the associations with syntactic frames and thematic grids. In practice we explored the best way of producing these associations by assessing the ability of the resulting classifications to support the labeling of verb

syntactic arguments with Verbnet thematic roles. This evaluation is presented in detail in Chapter 5, here we summarise the most prominent results.

Since the gold standard by [Sun *et al.*, 2010], which we used so far to compare our clustering with, does not provide associations with syntactic frames, in this evaluation we needed to use a different reference, namely the SRL gold standard introduced and described in Section 3.2.1.2. SRL gold is a reference corpus where verb syntactic arguments were manually labeled with Verbnet semantic roles.

Syntactic frames. For the association with syntactic frames, we used the **Fmax** and **Fpos** labeling schemes described in Section 3.3.2.2. We found the **Fpos** procedure to perform better. In addition high frequency frames (as for example the transitive frame) had to receive a specific treatment: since they are not discriminative they are not important for the clustering and therefore often don't have high feature f-measure.

Thematic grids. We found that the thematic grids associated with the verbs by the clustering induced a very low recall. We therefore opted for a different approach (the same as for associating FCA concepts with translated Verbnet classes) and retrieved for each verb cluster C_V, the Verbnet class on which this cluster had the highest F-measure (with respect to intersection of verb members, as for the FCA classification). The thematic grids associated with this class were then assigned to all verbs in C_V.

That is the thematic grid assigned to each cluster by IGNGF through cluster labeling were not used, only the verbs (to compare the cluster with each translated Verbnet classes). For syntactic frames however, we used the frames associated by the clustering with the verb cluster since subcategorisation frames are much more language specific (for instance, in French, prepositional objects split into three categories depending on whether they are introduced by *de*, à or any other prepositions) and using Verbnet syntactic frames would not support a direct match with a standard syntactic annotation of French. Our evaluation of the ⟨verb, frame⟩ associations in Chapter 5 supported this intuition.

To sum up, for the evaluation which will be described in detail in Chapter 4 and 5 we used IGNGF in the following configuration.

> The IGNGF clustering was built based on the **grid-scf-sem** feature set and by applying the IDF-norm weighting scheme. To determine the number of clusters, this number was varied such that the obtained partition reached an optimal F-measure and a number of clusters in the same order

of magnitude as the number of classes in the gold by [Sun *et al.*, 2010]. We thus obtained groupings of French verbs. In Chapter 5 we investigated how these verb groups are best associated with sets of subcategorisation frames and thematic role sets. The most basic configuration we started from is by using the **Fmax** scheme to label clusters with both subcategorisation frames and thematic role sets, that is clusters were associated with the subcategorisation frames and translated Verbnet classes which maximise the cluster Feature F-measure. The translated classes used are obtained with the **svm** method.

3.4 Conclusion

In this chapter we presented the resources we made use of to build our Verbnet like classifications of French verbs. We introduced the two techniques applied to create the classifications: the symbolic method called Formal Concept Analysis (FCA) and the neural clustering method Incremental Growing Neural Gas with Feature Maximisation (IGNGF). Both methods were used to classify the French verbs provided by available lexical resources. The classification was performed based on features which were also extracted from these lexical resources. We described the lexical resources and the extracted features. While the main features are subcategorisation frames provided by a subcategorisation lexicon built by merging existing French lexical resources, we also used additional syntactic and semantic features which were extracted from the same lexical resources for French. To provide the verb classes produced through the two clustering methods with thematic role sets we used translated Verbnet classes. A group of verbs is associated with the thematic role set of the Verbnet class with the largest proportion of shared verbs. The translated Verbnet classes were also used as features in the IGNGF clustering process: a verb was considered to have a thematic role set if it was a member of a translated Verbnet class with this role set. The translated Verbnet classes were produced in several ways, using available translation dictionaries.

We showed how we applied both classification methods to classify French verbs, using the feature sets described above: For FCA, we developed a concept filtering technique based on the sum of the concept stability and separation indices, for IGNGF we used the cluster labeling process to associate the groups of verbs with subcategorisation frame information. We performed a partial and preliminary qualitative discussion. It showed that the groups of verbs produced by both techniques were coherent: we found these classes to adequately describe the association between

verb sets, syntactic frames and semantic components.

We also introduced the two reference data sets which we used for evaluation: the gold standard proposed by [Sun *et al.*, 2010], called V-gold, and the reference called SRL gold, where corpus ⟨verb, syntactic argument⟩ instances were manually assigned Verbnet thematic roles. A detailed evaluation with respect to the produced groups of verbs and their association with frames and thematic grids for both methods, using these references, will be presented in Chapters 4 and 5. In this section we used the gold standard by [Sun *et al.*, 2010] to determine which feature set was most helpful for creating the verb classes. We experimented with groups of features combining subcategorisation frames with syntactic and/or semantic features. An interesting finding is, that semantic features helped to improve the classifications for both the FCA and the IGNGF method: For FCA the best performing feature set was **scf-sem**, combining subcategorisation frames with additional semantic features and for IGNGF the feature set **grid-scf-sem**, consisting of subcategorisation frames, additional semantic features and features derived from the translated Verbnet classes, performed best.

Finally, we determined which method for translating Verbnet classes yielded the "best" translated classes. We consider this to be the **svm** method for the following reasons. First, for this method, the distribution of verbs in the translated classes was closest to that of Verbnet, second, when using these translated classes with the FCA classification, the number of induced correct ⟨verb, thematic grid⟩ associations with respect to the gold was highest.

In the next chapters we evaluate the two classifications obtained using the FCA and IGNGF method with respect both to the induced groups of verbs, and to their associated syntactic and semantic information.

Chapter 4

Evaluating Semantic Verb Classes

Contents

This chapter is devoted to evaluating the two classification methods introduced earlier based on the groupings of verbs they produce. The evaluation consists in comparing the groups of verbs produced by the two classification techniques with the V-gold classification (proposed in [Sun *et al.*, 2010] and described in Section 3.2.1.1) in two ways. First, we used the same evaluation methodology as Sun *et al.*, namely by employing the clustering evaluation metrics modified purity and weighted class accuracy. Second, we associated the clusters or classes produced with translated Verbnet classes and thereby with Verbnet role sets. This step may have introduced noise, but it allowed us to associate the cluster/classes

with thematic grids and therefore to compare the produced classifications with the V-gold based on the ⟨verb, thematic grid⟩ associations they induce.

The chapter is structured as follows. We first introduce the evaluation metrics. We then describe our experimental setting and finally present the evaluation results.

4.1 Evaluation Metrics

We used two evaluation metrics to evaluate both our classifications. The first is modified cluster purity (mPUR), weighted class accuracy (ACC) and their associated F-measure. The second is recall, precision and F-measure of the ⟨verb, thematic grid⟩ pairs derived from the classification compared to those induced by V-gold. Since the clustering techniques differ considerably, we also assessed the obtained classifications using measures which are specific to each clustering method, as for example the verb and cluster coverage measures. A further measure, cumulated micro precision (CMP), which is only computed for IGNGF, could in principle be used for the FCA classification as well, but because of its recent introduction has not yet been applied to FCA. In the following we give a more detailed description of each of these metrics.

4.1.1 Cluster Purity, Accuracy and F-measure

To evaluate both our classifications with respect to the groups of verbs they generate we followed [Sun *et al.*, 2010] and used modified purity (mPUR), weighted class accuracy (ACC) and their F-measure. These were computed as follows. Each induced cluster C was assigned the gold class (its *prevalent class*, prev(C)) to which most of its member verbs belong. A verb is then said to be correct if the gold associates it with the prevalent class of the cluster it is in. Given this, purity is the ratio between the number of correct gold verbs in the clustering and the total number of gold verbs in the clustering[26]:

$$mPUR = \frac{\sum_{C \in \text{Clustering}, |\text{prev}(C)|>1} |\text{prev}(C) \cap C|}{\sum_{C \in \text{Gold}} \text{Verbs}_{\text{Clustering} \cap \text{C}}},$$

where $\text{Verbs}_{\text{C} \cap \text{Clustering}}$ is the number of verbs in the gold class C and in the clustering.

Accuracy represents the proportion of gold verbs in those clusters which are associated with a gold class, compared to all the gold verbs in the clustering. To

[26] Clusters for which the prevalent class has only one element are ignored

compute accuracy we associate to each gold class C_{Gold} a dominant cluster, ie. the cluster $\text{dom}(C_{\text{Gold}})$ which has most verbs in common with the gold class. Then accuracy is given by the following formula:

$$ACC = \frac{\sum_{C \in \text{Gold}} |\text{dom}(C) \cap C|}{\sum_{C \in \text{Gold}} \text{Verbs}_{\text{C}}}$$

Finally, F-measure is the harmonic mean of mPUR and ACC.

Note, that the mPUR and ACC metrics, as used in [Sun *et al.*, 2010] had to be adapted to be applicable to the FCA classification which, in contrast to the clusterings in [Sun *et al.*, 2010] and the IGNGF clustering which are crisp, is overlapping. The difference is that for a non-crisp clustering, the denominator is the total number of elements in the classes as opposed to simply the number of verbs for a crisp clustering.

4.1.2 Precision, Recall and F-measure

For a further evaluation, we exploited the implicit association of verbs with thematic grids in the V-gold. We identified each Levin class in the gold with a Verbnet class and its thematic role set (these associations are explicited in Table 3.9, first column, underlined). Thus effectively each verb in the gold was associated with one or more thematic grids and we could check how many of these associations were also produced by the classifications. For this, the generated verb clusters needed to be mapped to a thematic grid. This mapping was obtained by aligning the verb groups with the translated Verbnet classes: each cluster C_{classif} was associated with the Verbnet class C_{VN} with best f-measure between precision (P) and recall (R), computed as follows.

$$R = \frac{|C_{\text{classif}} \cap C_{\text{VN}}|}{|C_{\text{VN}}|}, P = \frac{|C_{\text{classif}} \cap C_{\text{VN}}|}{|C_{\text{classif}}|}.$$

Thus each cluster was effectively associated with a thematic grid and the resulting ⟨verb, grid⟩ pairs could be compared with those present in the gold: Precision is the proportion (*Classif*∩*Gold*)/*Classif* of verb/Verbnet class pairs found by our method that is correct. Recall is the proportion of ⟨gold verb, Verbnet class pairs⟩ that is found (*Classif*∩*Gold*)/*Gold*. And F-measure is the harmonic mean of precision and recall.

4.1.3 Metrics Applied to IGNGF Only

The following two metrics were applied only to the IGNGF clustering.

Cluster coverage. To assess the extent to which a clustering matches the gold classification, we additionally computed the *coverage* of each clustering that is, the proportion of gold classes that are prevalent classes in the clustering. This metric need not be applied to the FCA classifications because of the way the FCA concepts are associated with Verbnet classes. Recall that, for filtering the FCA concepts we first select the 1500 concepts where the sum of stability and separation is highest. Then, for each translated Verbnet class, we select from these 1500 concepts those where the proportion of common verbs and class size is best (see Section 3.3.1). Thus for FCA a maximal gold class coverage is warranted.

Cumulative Micro Precision (CMP). As pointed out in [Lamirel *et al.*, 2008; Attik *et al.*, 2006], unsupervised evaluation metrics based on cluster labeling and feature maximisation can prove very useful for identifying the best clustering strategy. Following [Lamirel *et al.*, 2011a], we used cumulative micro precision (CMP) to identify the best clustering. Computed on the clustering results and taking into account for each cluster, the (possibly weighted) features it contains, this metrics evaluates the quality of a clustering with respect to the cluster features rather than with respect to a gold standard. It can be applied on the overall clustering results and thus allows an evaluation based on the entire data space. This is complementary to an evaluation with respect to a gold standard, because in most cases the gold standard consists of a more or less representative selection from the objects to be classified. In [Ghribi *et al.*, 2010] CMP was shown to be effective in detecting degenerated clustering results including a small number of large heterogeneous, "garbage" clusters and a big number of small size "chunk" clusters. In the following we describe how the CMP score is derived.

First, the *local Recall* (R_c^f) and the *local Precision* (P_c^f) of a feature f in a cluster c are defined as follows:

$$R_c^f = \frac{|v_c^f|}{|V^f|}, \quad P_c^f = \frac{|v_c^f|}{|V_c|}$$

where v_c^f is the set of verbs having feature f in c, V_c the set of verbs in c and V^f, the set of verbs with feature f.

Based on these, we define the global clustering metrics *Macro-Recall* (MR) and *Macro-Precision* MP as follows:

$$MR = \frac{1}{|\overline{C}|} \sum_{c \in \overline{C}} \frac{1}{|F_c|} \sum_{f \in F_c} R_c^f \tag{4.1}$$

$$MP = \frac{1}{|\overline{C}|} \sum_{c \in \overline{C}} \frac{1}{|F_c|} \sum_{f \in F_c} P_c^f \tag{4.2}$$

where F_c is the set of maximising features for cluster c and $\overline{C}$ the set of clusters $c \in C$ for which F_c is non-empty.

As [Lamirel *et al.*, 2011a] shows, Macro Recall and Macro Precision can be seen as a measure of how much this clustering differs from the "exact" classification, namely the formal concept lattice built from the formal context given by the objects to be clustered and their features. If both Macro Precision and Macro Recall equal 1, $\overline{C}$ represents a sub-lattice of this concept lattice[27].

Macro recall and precision are cluster oriented measures, since they are computed based on the (local) precision and recall values for each cluster. They permit a global estimation of the optimal number of clusters (the one where macro recall and precision are balanced) but they fail to detect degenerated clusterings, which have a small number of large heterogeneous clusters and a big number of small clusters (which typically have high local precision). To capture these undesirable clusterings, [Lamirel *et al.*, 2004] propose the measures of *Micro Precision* and *Micro Recall*:

$$mR = \frac{1}{|V|} \sum_{c \in \overline{C}, f \in F_c} R_c^f \tag{4.3}$$

$$mP = \frac{1}{|V|} \sum_{c \in \overline{C}, f \in F_c} P_c^f \tag{4.4}$$

where V is the set of objects to be clustered. In contrast to macro precision and recall, micro precision and recall are feature oriented measures in that they average the local precision and recall disregarding the specific structure of the clusters. To combine the macro and micro precision measures and thus to capture the quality of the clustering both in terms of the structure of the clusters and of their maximising features, [Lamirel *et al.*, 2011a] introduce Cumulative Micro-Precision (CMP), which

[27] Considering that we built the FCA classification based on the same data, it is an interesting direction for future work to compare the two classifications based on these measures.

	mPUR & ACC	Precision/Recall/F	CMP	class coverage	verb coverage
FCA	yes	yes	no	=100%	yes
IGNGF	yes	yes	yes	yes	=100%

Table 4.1: Overview of evaluation metrics and for what clustering technique they are computed.

is defined as follows:

$$CMP = \frac{\sum_{i=|C_{inf}|..|C_{sup}|} \frac{1}{|C_{i+}|^2} \sum_{c \in C_{i+}, f \in F_c} P_c^f}{\sum_{i=|C_{inf}|..|C_{sup}|} \frac{1}{|C_{i+}|}} \tag{4.5}$$

where C_{i+} represents the subset of clusters of C for which the number of associated verbs is greater than i, and: $C_{inf} = argmin_{c_i \in C}|c_i|$, ie. the smallest cluster size, $C_{sup} = argmax_{c_i \in C}|c_i|$ ie. the largest cluster size. In this formula, a large weight is given to the large clusters, for which the probability to group incoherent data is higher. If the data in these clusters is indeed incoherent, this will lower the CMP value considerably, whereas large coherent clusters will have a positive influence on the CMP[28].

The following measure was computed only for FCA.

4.1.4 Metrics Applied to FCA Only

Verb coverage. Since when building the FCA classification we discarded a large number of concepts, we had to make sure that the selected concepts covered a large proportion of the verbs to be classified. Verb coverage is always 100% for IGNGF, because IGNGF always assigns all verbs to classes.

Table 4.1 gives an overview of the described evaluation metrics, for which classification they apply and for which they were computed.

4.2 Experimental Setting

Our verb classifications were obtained in two ways: First by the symbolic classification method FCA, as described in Section 3.3.1 and second by the neural clustering method IGNGF as shown in Section 3.3.1. We applied both clustering methods on

[28] In principle it would be possible to compute the CMP for FCA classifications as well, but we did not perform these computations yet.

the **vn_restricted** data set (Section 3.2.2) and used translated Verbnet classes obtained with the **svm** method (Section 3.1.2.1). The gold classification we compared our groups of verbs with was proposed in [Sun *et al.*, 2010] and was introduced in Section 3.2.1.1.

In the following we briefly recall for each technique how the classifications were built before presenting a comparative summary of the results in terms of the evaluation metrics introduced in Section 4.1.

4.2.1 FCA

First, as described in Section 3.3.1, we built a formal context based on the verbs in data set **vn_restricted** (Section 3.1.3) and the feature set **scf-sem**, which in Section 3.3 showed the best performance. From this formal context, FCA then constructs a concept lattice. According to the findings in Section 3.3.1.2, this concept lattice was filtered using the stability and separation concept selection indices: only the 1500 concepts with highest sum of concept stability and separation were selected. To associate these concepts with a thematic grid we aligned them with translated Verbnet classes. From the 1500 concepts we selected those which have the best precision-recall f-measure with respect to intersection with a translated Verbnet class. An excerpt of the resulting classification is shown in Figure 3.7.

4.2.2 IGNGF

We used IGNGF in the setting described in Section 3.3.2. The objects to be clustered were the verbs in data set **vn_restricted**. We used the feature set **grid-scf-sem** which was shown to perform best in the experiments in Section 3.3.2. We applied the IDF-norm weighting scheme described in Section 3.3.2. In order to better assess the performance of the IGNGF clustering method, we also compared it with a baseline computed using K-Means on the same verb and feature set. For each clustering method (K-Means and IGNGF), we let the number of clusters vary between 1 and 30 to obtain a partition that reaches an optimum F-measure and a number of clusters that is in the same order of magnitude as the initial number of Gold classes (i.e. 11 classes).

The verb clusters were associated with thematic grids by aligning them with the translated Verbnet classes: Each cluster was associated with the thematic grid of the translated Verbnet class with best F-measure with respect to the shared number of verbs.

(a) Purity, accuracy and F-measure for the classifications obtained with FCA and IGNGF, compared to gold classification in [Sun *et al.*, 2010].

	Purity	Accuracy	F-measure	CMP at opt. (13cl.)	Class coverage
FCA	32.30	95.61	48.29		100
IGNGF	86.00	59.00	70.00	0.30	72
K-Means	88.00	57.00	70.00	0.10	88
[Sun *et al.*, 2010]			55-65.4		

(b) Precision, recall and F-measure for ⟨verb, theta-grid⟩ pairs obtained with FCA and IGNGF classification.

	Verb coverage	Precision	Recall	F
FCA	96.17	24.09	75.00	36.47
IGNGF	100.00	27.16	26.67	27.16

Table 4.2: Evaluation of verb groups generated by FCA and IGNGF, based on modified purity, accuracy and their F-measure (a) and on a comparison of ⟨verb, theta-grid⟩ associations in the reference (b). The 4th column in (a) shows the cumulative micro precision index and the last column gives the number of classes with a prevalent **grid** feature.

4.3 Results - Comparative Summary

Table 4.2 shows the results for both classifications, based on purity, accuracy and their F-measure (Table 4.2a) and on a comparison of ⟨verb, theta-grid⟩ associations in the reference.

4.3.1 Evaluating Groups of Verbs

We first comment on the purity, accuracy and F-measure results (Table 4.2a). First we note that the IGNGF F-measure outperforms [Sun *et al.*, 2010] whose best F-measures vary between 55 for verbs occuring at least 150 times in the training data and 65 for verbs occurring at least 4000 times in this training data. The results are not directly comparable however since the gold data is slightly different due to the grouping of Verbnet classes through their thematic grids. More importantly, our features were acquired from lexical resources, whereas those of [Sun *et al.*, 2010] were automatically extracted from a large corpus. Interestingly, despite these differences, [Sun *et al.*, 2010] also report an increase in performance when using semantic features, an observation confirmed for both our classification methods.

For the IGNGF clusters, the unsupervised clustering metrics indicated strong cluster cohesion with a number of clusters close to the number of gold classes (5th column in Table 4.2a), 13 clusters for 11 gold classes) and a high Cumulated Micro Precision (CMP = 0.30 vs. 0.10 for K-means). Although K-means and IGNGF reach similar F-measure and display a similar number of clusters, the very low CMP (0.10) of the K-means model shows that, despite a good gold class coverage, K-means tends

to produce more large and incoherent clusters. The coverage of 72% indicates that approximately 8 out of the 11 gold classes could be matched to a prevalent gold class.

Comparing IGNGF with FCA, the Table 4.2a shows that the IGNGF classification achieved much better results with respect to the global clustering evaluation metrics than the FCA classification (roughly 20 points difference in F-measure). Interestingly, for IGNGF purity was higher than accuracy (86 and 59 respectively) whereas the FCA classification had a very high accuracy and a much lower purity (95.61 and 32.30 respectively). Recall that purity is the proportion of correct verbs in the clustering with respect to the total number of gold verbs in the clustering, whereas accuracy represents the proportion of gold verbs in the clusters associated with a gold class compared to the total number of gold verbs in the clustering. Thus, the high purity of the IGNGF clustering suggests that an important part of the verbs are grouped in a similar way as in the gold, whereas the low purity of the FCA classification indicates an important structural difference to the gold. This is not too surprising considering the overlapping nature of the FCA classification. On the other hand accuracy results suggest that a large proportion of gold verbs are grouped in a similar way in the FCA classification as in the gold, which appears to be less true for the IGNGF clustering.

4.3.2 Evaluating ⟨verb, thematic grid⟩ Associations.

In contrast, when evaluating the induced ⟨verb, theta-grid⟩ pairs, the FCA classification outperformed the IGNGF classification by about 10 points in F-measure. In particular, whilst for IGNGF recall and precision are similar, for FCA precision is about 3% lower but recall is very much higher (50%). The reason for this is the overlapping nature of the FCA classification, due to which a verb may be associated to several thematic grids. The low precision indicates that many ⟨verb, theta-grid⟩ associations produced by the FCA or IGNGF classifications were not correct, according to the gold. However, some of these associations may be correct but just not present in the gold. For example in the gold the verb *déprimer* is a member of the *Cause-AgExp* (amuse) class only, whereas in the manually annotated reference in Chapter 3.2.1.2, it only occurred in the sense of *decrease*, ie. the *AgExp-Instrument-Patient* class. The FCA classification correctly produced both these associations, while IGNGF produced only ⟨*déprimer*, *Cause-AgExp*⟩.

To sum up, this evaluation shows that both the FCA and the IGNGF classifications produced coherent groups of verbs. This coherence is more evident for the IGNGF method, where the produced clustering, compared to the gold standard, resulted in better purity/accuracy F-measures than related work presented in [Sun *et*

al., 2010]. This evidence is somewhat weakened by the fact that the results in [Sun *et al.*, 2010] are not entirely comparable to ours. On the other hand it is supported by the high CMP score for IGNGF.

The coherence of the FCA classes is supported by the high recall (75%) of the derived ⟨verb, thematic grid⟩ pairs with respect to those induced from the gold standard. The resulting F-measure of 36.47 is low due to the low precision (of 24.09%) which indicates that many ⟨verb, thematic grid⟩ pairs induced by the FCA classes are not present in the gold standard. This is however not necessarily wrong because the gold standard does not contain all possible ⟨verb, thematic grid⟩ pairs. Nevertheless, in terms of F-measure, the FCA classification outperforms the IGNGF clustering. One must consider however, that the comparison with the gold standard based on the associations of verbs and thematic grids relies on the associations of groups of French verbs with translated Verbnet classes, which possibly introduce noise.

Chapter 5

Evaluating Syntactic-Semantic Verb Classes

Contents

In Chapter 4 we evaluated the groups of French verbs generated by our classification methods by comparing them to the gold classification proposed in [Sun *et al.*, 2010]. Since this gold classification only assigns French verbs to Levin style classes,

this comparison assessed the semantic cohesion of the classes, but did not permit an evaluation of the associations of verbs with syntactic frames produced with our classifications.

> This chapter presents an evaluation of the classifications acquired using Formal Concept Analysis – FCA (as described in Section 3.3.1) and Incremental Growth Neural Gas with Feature maximisation clustering – IGNGF (Section 3.3.2) with respect not only to the groups of verbs but also the subcategorisation frames and thematic grids associated with each verb by these classifications. We assess the ability of the classifications to associate verbs with appropriate subcategorisation frames and thematic grids by comparing these associations with those provided by the reference corpus SRL gold, where verb arguments have been manually labeled with thematic roles, as described in Section 3.2.1.2. We perform this evaluation in two ways: First, on a global level, we compare the ⟨verb, frame⟩ and ⟨verb, grid⟩ pairs in the reference with those generated by the classifications. Second, we perform a task based evaluation by using the classifications to label the syntactic arguments of verbs with thematic roles. We thus assess the ability of the acquired classifications to support the correct assignment of thematic roles to syntactic arguments and provide a more fine grained evaluation of ⟨verb, syntactic argument⟩ associations on the syntax/semantics interface level.

The chapter is structured as follows. We start by describing the experimental setting for this evaluation in Section 5.1. We specify the data and how the two classification methods are used to build Verbnet like classifications from this data. In Section 5.2 we explain our evaluation methodology: On a global level we assess the ability of our classifications to induce the associations of verb, frames and thematic grids present in the reference corpus. On the syntax/semantics interface level, we perform a task based evaluation by assessing the support of the acquired classification in the assignment of thematic roles to ⟨verb, syntactic argument⟩ instances. Sections 5.2.2 and 5.2.3 show how the two classifications, obtained with the two classification methodologies FCA and IGNGF, perform with respect to each of the two evaluation strategies. Finally, Section 5.2.4 presents a comparative summary and Section 5.3 summarises the insights and directions for further work which we draw from this evaluation.

5.1 Experimental Setting

In the following experiments we use FCA and IGNGF based on data set **vn_all**, described in Section 3.2. That is the verbs to be clustered were all verbs present in the merged subcategorisation lexicon. In essence the feature sets used for clustering/classification by the two methods were the same (feature set **scf-sem**), namely the subcategorisation frames associated to the verbs by the lexicon (shown in Section 3.1.1) and the additional semantic features shown in Table 3.5b. For IGNGF we used the feature set **grid-scf-sem**: In addition to the features in **scf-sem**, we used the **grid** features derived from the translated Verbnet classes: a verb "had" a thematic role set, if it was in a translated Verbnet class with that thematic grid. As before, the translated Verbnet classes were obtained using the **svm** method (Section 3.1.2.1).

5.1.1 FCA

Based on the **vn_all** data we built a formal context consisting of 4260 objects (verbs) and 303 attributes (SCFs and semantic features). The concept lattice resulting from this formal concept had 35274 concepts. These were filtered as described in Section 3.3.1.2 using the sum of the stability and separation indices. Following the procedure described in Section 3.3.1.3, these concepts grouping verbs and subcategorisation frames were associated with thematic role sets by aligning the concepts with translated Verbnet classes based on the verbs they have in common. Only those concepts were kept in the final classification which could be assigned a thematic role set. The resulting FCA classification groups 3994 verbs into 52 classes associated with a total of 32 subcategorisation frames and 61 thematic role sets.

Figure 5.1 shows an excerpt of the obtained classification. For example class 9109 contains 59 verbs (*eg. abaisser, accompagner, acheminer, apporter, avancer, baisser, balancer, bouger, cahoter, camionner, catapulter, charrier, colporter, coltiner, ...*) which can all be used in the constructions given by the subcategorisation frames SUJ:NP,OBJ:NP (basic transitive construction) and SUJ:NP,OBJ:NP,POPJ:PP,-POBJ:PP (transitive construction with two additional prepositional objects). Semantically it is associated with the set of semantic roles Agent-End-Start-Theme. The Verbnet classes with this set of thematic roles are: *send-11.1, carry-11.4, drive-11.5, banish-10.2, slide-11.2, bring-11.3.*

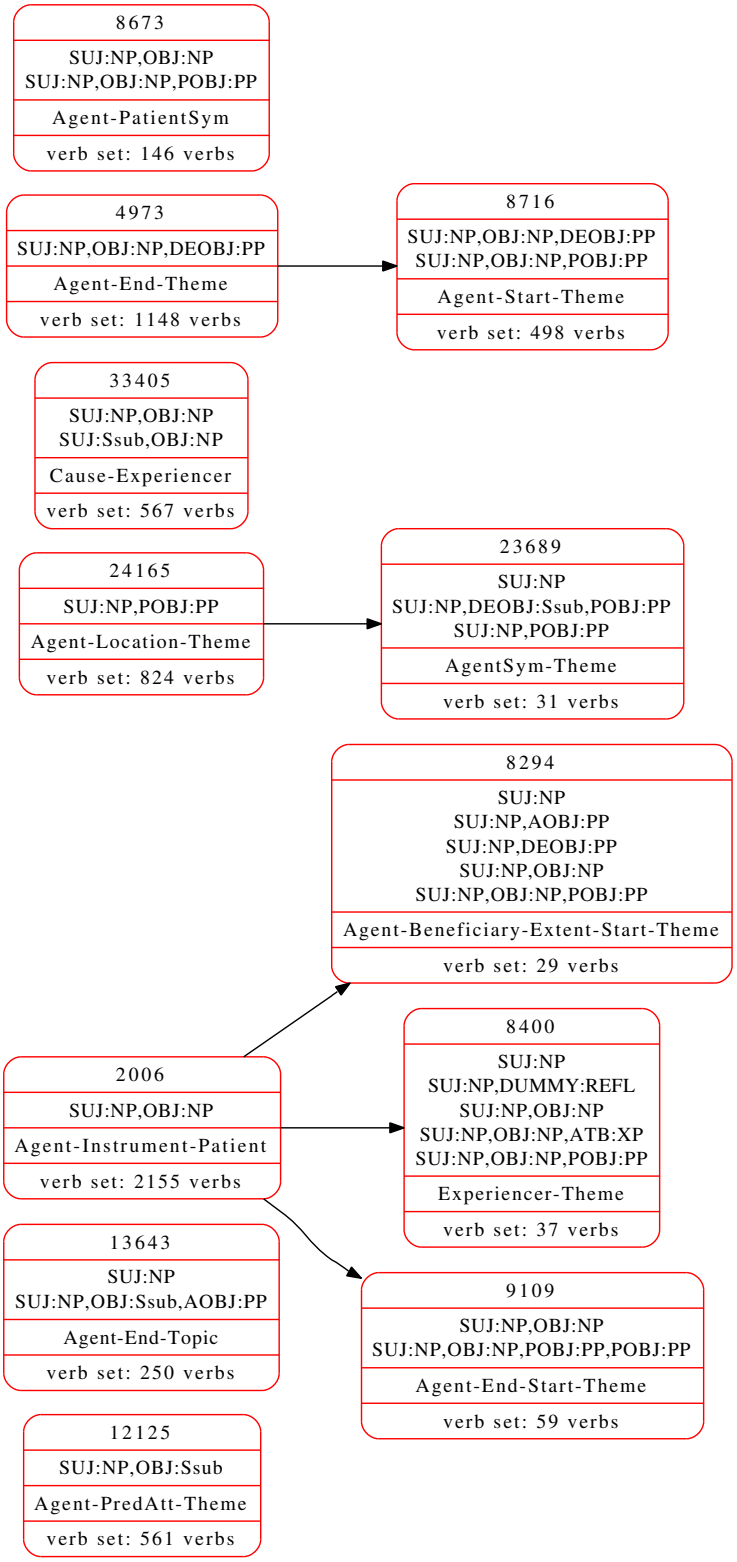


Figure 5.1: Excerpt of the verb classes obtained using FCA.

5.1.2 IGNGF

We applied the IGNGF clustering method on the verbs in data set **vn_all** and thus obtained clusters of verbs. To obtain a Verbnet like classification, these clusters needed to be associated with subcategorisation frames and with thematic role sets. To associate the clusters with subcategorisation frames and thematic grids, we used the cluster labeling techniques described in Section 3.3.2.2. Since both the syntactic frames and thematic grids were among the features used they could in principle be used to "label" the clusters and thus provide the associations with syntactic frames and thematic grids. However, as we will show in Section 5.2.3, we found that the thematic grids associated by the clustering with the verb clusters induced a very low recall. We therefore opted for a different approach and retrieved for each verb cluster C_V, the Verbnet class on which this cluster had the highest F-measure (with respect to intersection of verb members). The thematic grids associated with this class were then assigned to all verbs in C_V. That is, in this case, the thematic grids assigned to each cluster by IGNGF were not used, only the verbs (to compare the cluster with each translated Verbnet class). For syntactic frames however, we used the frames associated by the clustering with the verb cluster since subcategorisation frames are much more language specific (for instance, in French, prepositional objects split into three categories depending on whether they are introduced by *de*, à or any other prepositions) and using Verbnet syntactic frames would not support a direct match with a standard syntactic annotation of French.

In sum, while we relied on Verbnet and on the verb clusters to acquire thematic grids, we retrieved syntactic frames from the cluster labeling provided by IGNGF. The evaluation on thematic grids thus yielded additional insights on the quality of the verb clustering whilst the results on syntactic frames shed light on the value of the feature profiles associated by IGNGF with each cluster. The process of finding the best way to label the clusters with syntactic frames and thematic grids is further detailed in Section 5.2.3.

5.2 Evaluation

As stated earlier, the associations between verbs, subcategorisation frames and thematic grids, provided by the two types of classifications, were assessed on two levels. The first is a global level, where we looked at how well the acquired classifications associated verbs with subcategorisation frames and thematic grids present in the SRL gold corpus. The second level was concerned with the syntax/semantics interface. Here we checked whether and to what extent the acquired classifications support the

correct assignment of thematic roles to syntactic arguments – in essence a simplified semantic role labeling task.

In the following, we first give more details on how each of these evaluations is conducted. Then we analyse the FCA and IGNGF classifications with respect to both these evaluation criteria. Finally we compare the two classification methods on both evaluation levels.

5.2.1 Methodology

5.2.1.1 Global level evaluation.

As mentioned above, for the global level evaluation (association of verbs with thematic grids and syntactic frames), we simply compared the frames and grids present in the SRL gold corpus with those given by the classifications.

5.2.1.2 Syntax/semantics interface level.

For the assessment of the syntax-semantic associations between syntactic arguments and semantic roles supported by the classification we proceeded as follows. We used each classification to label the ⟨verb, syntactic argument⟩ pairs in the reference with thematic roles and compared these associations to those given by the reference (SRL gold). To associate the ⟨verb, syntactic argument⟩ pairs with thematic roles, we used the procedure presented in [Swier and Stevenson, 2005], where Verbnet is used to bootstrap a semantic role labeler for English.

In the following we describe the method used by Swier and Stevenson and our adaptation to French and the acquired classifications. We then show how we used the resulting procedure to better assess the ability of our classifications to associate verbs and their syntactic arguments with Verbnet thematic roles.

SRL for English: The method of [Swier and Stevenson, 2005]. Swier and Stevenson, [2004] and [2005] present an unsupervised probabilistic semantic role labeling system for English that relies on the Verbnet lexicon (and its semantic role inventory) combined with a simple probabilistic method. The role labeling is performed on sentences in FrameNet II which are taken from the British National Corpus [Clear, 1993]. The authors chose this corpus because, since Verbnet has no associated semantic role labeled corpus, there is no labeled reference data available against which to evaluate the results of a Verbnet role labeling system. They obtained such a reference from the FrameNet corpus, by mapping FrameNet's larger set of semantic roles to Verbnet's much smaller one.

theta-grids for V	syntactic construction		%θ	%SCF	Score
	SUBJ	OBJ			
Agent V	Agent		100	50	150
Agent V Theme	**Agent**	**Theme**	100	100	**200**
Instrument V Theme	**Instrument**	**Theme**	100	100	**200**
Agent V Recipient Theme	Agent	Theme	67	100	167

Table 5.1: An example of frame matching. The syntactic arguments to be labeled are SUBJ and OBJ (for some verb V). According to the Verbnet classes V is a member of, V may occur with the theta grids listed in the leftmost column. The theta grids are scored with the proportion of semantic roles which can be aligned with a syntactic argument (%θ) plus the proportion of syntactic arguments which can be aligned with a semantic role (%SCF).

The FrameNet sentences were parsed, resulting in associations of verbs with automatically extracted syntactic frames of the form SUBJ V OBJ. Using Verbnet and a **frame matching** procedure, Swier and Stevenson processed these ⟨verb, frame⟩ pairs to detect those ⟨verb, syntactic argument, thematic role⟩ associations which are most probably unambiguous. The frame matching procedure is applied to a ⟨verb, frame⟩ pair and works as follows. The target verb in the pair is a member of one or more Verbnet classes. Each Verbnet class gives the set of thematic grids valid for verbs in this class. The frame in the pair is aligned with each of these thematic grids and each thematic grid receives a score indicating how well the roles in the thematic grid can be matched to the syntactic arguments in the frame. The score is the proportion of syntactic arguments in the frame which can be aligned with a thematic role in the thematic grid plus the proportion of thematic roles in the thematic grid which can be aligned with a syntactic argument in the frame.

Table 5.1 illustrates the procedure. The syntactic frame to be aligned with thematic grids is SUBJ V OBJ. The Verbnet verb class(es) this verb is a member of show that the verb may occur with the theta-grids listed in the leftmost column. The theta-grids **Agent V Theme** and **Instrument V Theme** scored best and are selected. These two selected theta-grids are aligned with the SUBJ V OBJ subcategorisation frame and from this alignment SUBJ is associated with the semantic roles **Agent** and **Instrument** and OBJ with **Theme**. As only the OBJ $\leftrightarrow$ **Theme** association is unambiguous, the ⟨verb, OBJ⟩ instances in SUBJ verb OBJ constructions in the corpus sentences are labeled with the semantic role **Theme**.

In the next step the resulting unambiguous ⟨verb, syntactic argument, semantic role⟩ associations are used to compute probability estimates

$$P(\text{semantic role}|\text{syntactic argument class})$$

where the syntactic argument classes are those used in chunking, disregarding prepo-

SRL method	F-measure
baseline	0.74
frame matcher	0.76
frame matcher and probability estimates	0.83

Table 5.2: SRL results in [Swier and Stevenson, 2005].

sitions: subject, object, indirect object and prepositional object. These probability estimates are used to disambiguate the remaining ambiguous ⟨verb, syntactic argument, semantic role⟩ associations. That is, a ⟨verb, syntactic argument⟩ pair is associated with the semantic role with highest probability estimate

$$P(\text{semantic role}|\text{syntactic argument}).$$

For evaluation, Swier and Stevenson compare the obtained ⟨verb, syntactic argument class, semantic role⟩ associations to those in a reference. The reference consists of 1000 sentences from the FrameNet corpus, where the syntactic arguments identified by the parser and the FrameNet to Verbnet role mappings were manually corrected, resulting in ⟨verb, syntactic argument⟩ pairs labeled with Verbnet semantic roles. The baseline consists in assigning to each ⟨verb, syntactic argument⟩ pair the semantic role most frequently associated with this syntactic argument in the corpus sentences used as development set. Results are shown in Table 5.2. The authors only report the F-measure stating that precision and recall were always close in value.

Using Swier and Stevenson's method to evaluate verb classifications. In the following we show how we adapted Swier and Stevenson's method described in the previous section to be applicable to our data for French. The English resources involved in Swier and Stevenson's method are Verbnet and the English ⟨verb, syntactic argument⟩ instances derived from the corpus. In our adaptation we used our acquired classifications instead of Verbnet and ⟨verb, syntactic argument⟩ instances extracted from the French SRL gold corpus instead of the English ⟨verb, syntactic argument⟩ pairs. The comparison of the produced ⟨verb, syntactic argument, thematic role⟩ associations with the reference annotations in SRL gold then allowed to assess the support of our verb classifications to Swier and Stevenson's method for assigning thematic roles to ⟨verb, syntactic argument⟩ instances.

The global outline of the adapted method is the same as in [Swier and Stevenson, 2005]: As Swier and Stevenson, we first built a seed labeling from our classifications of French verbs by extracting unambiguous ⟨verb, syntactic argument,

6583: 50 verbs, 5 frames, 1 role set	
frames	SUJ:NP SUJ:NP,OBJ:NP SUJ:NP,OBJ:NP,AOBJ:PP SUJ:NP,OBJ:NP,DEOBJ:PP SUJ:NP,OBJ:NP,DUMMY:REFL
thematic role set	Agent-Beneficiary-Start-Theme (VN *steal-10.5* class)
verbs	acheter, assurer, attendre, attirer, avancer, câbler, connaître, cracher, croire, découvrir, demander, défendre, dire, délivrer, donner, envier, faire, faucher, fermer, fixer, flanquer, fournir, gagner, indiquer, inspirer, jeter, masquer, passer, payer, piquer, porter, prendre, présenter, ramener, rappeler, rapporter, reconnaître, rejeter, rendre, reprendre, représenter, servir, sortir, souffler, tendre, tenir, tirer, trouver, vendre, voler

Table 5.3: Sample French Verbnet like class.

semantic role⟩ associations. We then used these associations to label ⟨verb, syntactic argument⟩ instances in the SRL gold corpus with thematic roles. From this labeling we computed probability estimates P(semantic role|syntactic argument). Then a ⟨verb, syntactic argument⟩ pair in the SRL gold corpus, which had not yet been assigned a semantic role, was labeled with the semantic role with highest P(semantic role|syntactic argument). We now describe in more detail our adaptation of Swier and Stevenson's frame matching procedure to French.

Frame matching. As just mentioned, we used our acquired classifications instead of Verbnet. A sample French verb class produced by one of these classifications[29] is shown in Table 5.3. To apply the frame matching presented in [Swier and Stevenson, 2004], we need to associate our verb classes with thematic grids. This association is achieved as follows: Our classes are associated with Verbnet thematic role sets (Agent-Beneficiary-Start-Theme for the example in Table 5.3) so we could extract from Verbnet the classes with these thematic roles. These Verbnet classes provide (among other things) the thematic grids these thematic roles can occur in and we could associate our verb classes with these thematic grids. For example, class 6583 shown in Table 5.3 is associated with the role set Agent, Beneficiary, Start, Theme. The Verbnet class with this role set is the class *steal-10.5*[30]. Verbnet lists the following theta-grids for the *steal-10.5* class:

Agent V Theme
Agent V Theme {source preposition} Start
Agent V Theme {for} Beneficiary
Agent V Theme {source preposition} Start {for} Beneficiary

[29] FCA in this case
[30] In this case there is only one Verbnet class with this role set, but there may be more

Class ids	SCFs	Role sets (Verbnet classes)
12119	SUJ:NP,OBJ:NP	Agent-Beneficiary-End-Start
	SUJ:NP,OBJ:Ssub	Cause-End-Topic
12120	SUJ:NP,OBJ:Ssub	Agent-PredAtt-Theme
8721	SUJ:NP,OBJ:NP,POBJ:PP	Agent-ThemeSym
8718	SUJ:NP,OBJ:NP	Agent-Extent-Start-Theme
	SUJ:NP,OBJ:NP,POBJ:PP	
8720	SUJ:NP,OBJ:NP	Agent-Beneficiary-End
	SUJ:NP,OBJ:NP,POBJ:PP	Agent-End-Location-ThemeSym
2001	SUJ:NP,OBJ:NP	Agent-Start-Topic
	SUJ:NP	
44	SUJ:NP	Agent-Beneficiary-End-Extent-Start
		Agent-Extent-Location
		Agent-PredAtt
2003	SUJ:NP,OBJ:NP	Agent-Instrument-Patient

Table 5.4: Classes with associated subcategorisation frames and role sets.

We drop prepositions and omit the position of the verb since we don't expect them to be the same in French, so the theta-grids we would finally use for our method in this example are Agent-Theme, Agent-Theme-Start, Agent-Theme-Beneficiary and Agent-Theme-Start-Beneficiary. As for English, subcategorisation frames were aligned with the Verbnet theta grids. However, here the subcategorisation frames were provided by the verb classes and thus ultimately come from the lexicon, whereas for English they were obtained by chunking. We assumed that the order of the semantic roles in the theta-grids in French and English are similar, whereas we expected the syntactic arguments to be different. We thus built ⟨verb, syntactic argument, semantic role⟩ associations for each verb in a class and each of the syntactic arguments in a subcategorisation frame associated to that class.

In the following we illustrate the way the adapted frame matching works by considering the example of the verb *concevoir* and a classification obtained by FCA. The verb *concevoir* is in the FCA classes 12119, 12120, 8721, 8718, 8720, 2001, 44, 2003. The associations to frames and role sets for these classes is shown in Table 5.4.

Table 5.5 shows the scores for verb *concevoir*, the syntactic frame SUJ:NP,-OBJ:NP,POBJ:PP and the theta-grids associated to this verb and subcategorisation frame by our classification. *concevoir* is associated with the syntactic frame SUJ:NP,-OBJ:NP,POBJ:PP by the FCA classes shown in the leftmost column. These FCA classes (8721, 8718 and 8720) were associated by our method with the thematic role sets (Verbnet classes) shown in the leftmost column in Table 5.5 and in the right column in Table 5.4. We retrieved from these (English) Verbnet classes the thematic grids which show how the thematic roles of a particular class are realised in syntac-

FCA class role set	theta-grids for *concevoir*	syntactic construction			%θ	%SCF	Score
		SUJ:NP	OBJ:NP	POBJ:PP			
8721	Agent-Theme	Agent	Theme		100	67	167
Agent,	Agent-Theme	Agent	Theme		100	67	167
ThemeSym	Theme-Theme	Theme	Theme		100	67	167
	Agent-Theme-Theme	**Agent**	**Theme**	**Theme**	100	100	**200**
8718	Agent-Theme	Agent	Theme		100	67	167
Agent, Extent	Extent-Theme	Extent	Theme		100	67	167
Start, Theme	**Agent-Theme-Start**	**Agent**	**Theme**	**Start**	100	100	**200**
	Agent-Theme-Extent	**Agent**	**Theme**	**Extent**	100	100	**200**
8720	Agent-End	Agent	End		100	67	167
Agent, Benef,	**Agent-Benef-End**	**Agent**	**Benef**	**End**	100	100	**200**
End	**Agent-End-Benef**	**Agent**	**End**	**Benef**	100	100	**200**
Agent, End	Agent-Theme	Agent	Theme		100	67	167
Location,	Theme-Location	Theme	Location		100	67	167
ThemeSym	Theme-Theme	Theme	Theme		100	67	167
	Agent-Theme-End	**Agent**	**Theme**	**End**	100	100	**200**
	Agent-Theme-Theme	**Agent**	**Theme**	**Theme**	100	100	**200**
	Theme-Location-Theme	**Theme**	**Location**	**Theme**	100	100	**200**

Table 5.5: An example of frame matching for French. The syntactic arguments to be labeled are SUJ:NP, OBJ:NP and POBJ:PP for *concevoir*. The FCA classes *concevoir* is a member of (displayed in leftmost column), are associated with the Verbnet thematic grids listed in the second column. The theta grids are scored with the proportion of semantic roles which can be aligned with a syntactic argument (%θ) plus the proportion of syntactic arguments which can be aligned with a semantic role (%SCF). In bold face the best scoring theta grids.

tic constructions. These thematic grids, shown in the second column of Table 5.5 are aligned with the syntactic frame SUJ:NP,OBJ:NP,POBJ:PP. They are scored according to a heuristic reflecting how well the syntactic arguments and thematic roles could be matched. The 8 thematic grids achieving best score (200) are set in bold face. When aligning these grids to the syntactic arguments in the frame the thematic role $\leftrightarrow$ syntactic argument associations were as follows:

	SUJ:NP	Agent, Theme
concevoir	OBJ:NP	End, Beneficiary, Theme, Location
	POBJ:PP	Beneficiary, End, Extent, Start, Theme

Since we only kept the unambiguous associations as label candidates and none of the above associations was unambiguous, none was kept in the final labeling. This example shows that this frame matching method is very restrictive.

In the following we explore a less restrictive frame matching method. Each subcategorisation frame can be mapped either to all the theta-grids associated to the verb or to those theta-grids associated to classes where the verb has the subcategorisation frame under consideration. The former method is more restrictive because a single subcategorisation frame is mapped to more theta-grids and thus the syntactic argument $\leftrightarrow$ semantic role associations are more likely to be ambiguous. Let us consider

FCA class	syntactic construction		
	SUJ:NP	OBJ:NP	POBJ:PP
8721	**Agent**	**Theme**	**Theme**
8718	**Agent**	**Theme**	
8720			

Table 5.6: Unambiguous role assignments with the less restrictive frame matching method. The setting is the same as in Table 5.5, but the frame matching is performed separately for each FCA class containing *concevoir* and associated with the syntactic frame SUJ:NP,OBJ:NP,POBJ:PP. We list, for each such FCA class the syntactic arguments which could unambiguously be associated with a thematic role.

again the case of ⟨*concevoir*, SUJ:NP,OBJ:NP,POBJ:PP⟩. With the less restrictive frame matching method, the syntactic frame would be matched *separately* with the theta-grids of each of the classes which are associated with this frame: 8718, 8720, 8721. The obtained unambiguous associations are shown in Table 5.6. We see that with the less restrictive frame matching method, the following three ⟨verb, syntactic argument⟩ pairs could be unambiguously associated with thematic roles: ⟨*concevoir*, SUJ:NP⟩, ⟨*concevoir*, OBJ:NP⟩ and ⟨*concevoir*, POBJ:PP⟩ could unambiguously be assigned the thematic roles Agent, Theme and Theme respectively. In contrast, as we saw earlier, with the more restrictive frame matching method, none of the syntactic arguments could be unambiguously labeled with a thematic role.

In the following we experimented with both the more and less restrictive frame matching method. However, as the IGNGF clusterings are crisp, ie. non-overlapping, the restrictive and non-restrictive frame matching produce the same results for this method.

Combining frame matching with probability estimates. As in [Swier and Stevenson, 2005], the unambiguous ⟨verb, syntactic argument, thematic role⟩ associations obtained by frame matching were first used to label ⟨verb, syntactic argument⟩ instances in the SRL gold. From these labeled corpus instances we then computed probability estimates P(thematic role|syntactic argument), ie. we counted the number of times each thematic role is associated to each syntactic argument.

Using Swier and Stevenson's SRL method for evaluation. For our evaluations we used each of our verb classifications to build two SRL systems, which label ⟨verb, syntactic argument⟩ corpus instances with Verbnet thematic roles. The first was based on frame matching only and thus only made use of our classifications built from lexical resources. Comparing the labelings obtained by the systems built this way with the reference data shows to what extent the classifications can account

for the associations of verbs, syntactic argument and thematic roles present in the reference corpus. The second type of SRL system combines frame matching with probability estimates extracted from corpus data. A comparison of the resulting labelings with the reference data gives information on how well the associations obtained with frame matching can be adjusted or complemented by corpus data. The produced labelings were compared with the reference annotations on one hand and default associations on the other. For the default associations, syntactic arguments were associated with thematic roles as follows:

SUJ:NP	Agent
OBJ:NP OBJ:Ssub OBJ:VPinf	Theme
AOBJ:PP AOBJ:VPinf	End
DEOBJ:PP DEOBJ:VPinf	Start
POBJ:PP POBJ:Ssub POBJ:VPinf	Location
ATB:XP	PredAtt

In the following we first present the results for the FCA and IGNGF classifications separately before showing a more general comparison.

5.2.2 The FCA Classification

In this section we discuss the evaluation results obtained when using the FCA classification built as explained in Section 5.1.1. We first performed an evaluation on a global level in Section 5.2.2.1, where we evaluated the associations of verbs with subcategorisation frames and thematic grids. A more fine grained assessment of the associations of ⟨verb, syntactic argument⟩ pairs with thematic roles supported by the FCA classification is conducted in Section 5.2.2.2.

5.2.2.1 Global level evaluation.

For the global level evaluation (association of verbs with thematic grids and syntactic frames), we compared the frames and grids present in the reference annotations with those given by the classification.

Table 5.7 summarises the comparison of the associations between verbs, frames and thematic grids yielded by the FCA classification and those induced from the SRL

(a) Associations between verbs and **frames**.

SRL gold	classif	SRL gold and lex	SRL gold and classif	SRL gold, lex, not classif	SRL gold not lex	Recall	Recall w/o missing in lex
316	16542	274	243	31	42	76.90	88.69

(b) Associations between verbs and **theta grids**.

SRL gold	classif	SRL gold and classif	Recall
318	33525	280	88.05

Table 5.7: Global level evaluation of the association between verbs and syntactic frames (a) and verb and thematic grids (b) provided by the FCA classification (*classif*). The associations with frames are compared with the reference annotations (*SRL gold*) and the lexicon (*lex*) and the associations with thematic grids are compared with the associations in the reference annotations. The table shows types, ie. the number of distinct ⟨verb, frame⟩ or ⟨verb, thematic grid⟩ pairs.

gold corpus (syntactic frames and thematic grids) and the lexicon (only syntactic frames).

We first focus on the ⟨verb, frame⟩ associations shown in Table 5.7a. The recall for frames was 76.90% for types (84.12% for tokens, not shown in the table). This score is however distorted by the 42 (13.92%) ⟨verb, frame⟩ associations in the reference annotations which are not present in the lexicon and therefore cannot appear in the classification. When only considering the ⟨verb, frame⟩ pairs present in the lexicon and the reference, the recall is of 88.68%. Out of the 274 ⟨verb, frame⟩ pairs in the reference and the lexicon, 31 (11.31%) were not yielded by the classification, ie. the concepts they are members of were discarded in the filtering process. These associations are listed in Table B.1 in Appendix B. The frames which were missed most often are SUJ:NP,AOBJ:PP (occuring in 14 pairs) and SUJ:NP,OBJ:NP,ATB:XP (occuring in 5 pairs).

This indicates that the produced classification misses many relevant ⟨verb, frame⟩ associations. Looking at the classes, we found that in many cases the groups of verbs were associated with only one or two frames: 9 classes have one frame only and 21 have two frames. Moreover, the associated frames are often high frequency frames (SUJ:NP,OBJ:NP and SUJ:NP)[31] which don't help to distinguish one group of verbs from another. This suggests that, to improve the classification, we need to focus more on the associated frames. One way to achieve this is to integrate the attribute concept stability index into the filtering process, which measures a concept's coherence with respect to frames [32]. This could help to select concepts with the missing frames

[31] One class has only the transitive frame and another one only the intransitive frame. Of the 21 classes with two frames, 16 also contain one of the transitive or intransitive frame.

[32] Recall that we currently use the object concept stability index and the separation index for filtering. The object concept stability index measures the concept's coherence with respect to verbs

and to bring forward more coherent frame groups. A second possible direction for improvement is to further constrain the selection of concepts based on their syntactic frame set: One could for example require that to be selected a concept needs to have at least two syntactic frames and at least one of these be different from the transitive (SUJ:NP,OBJ:NP) and intransitive (SUJ:NP) frame.

Because the reference only contains a limited number of sentences, there are much more frames in the classification than in the reference, hence we did not compute precision.

Regarding now the thematic grids, of the 318 ⟨verb, grid⟩ pairs present in the reference, 280 (88.05%) were compatible with a corresponding ⟨verb, thematic role set⟩ pair generated by the classification, that is the roles in the corpus grid also occurred in a role set associated with this verb by the classification. 38 pairs present in the reference, almost 12%, were not produced by the classification. These associations are listed in Table B.2 in Appendix B. Analysing these 38 pairs (corresponding to 29 verbs) we found the following potential reasons:

Annotation issues. In 10 cases there were annotation issues: In 8 cases the annotation was wrong and in the other two cases an annotation with a different role set was conceivable. In 8 of the 10 cases the corrected ⟨verb, thematic grid⟩ association was compatible with the FCA association.

Semantic role confusions. In 5 cases the role in the annotation was related to the role provided by the classification. A typical example is the Location role vs. the Source (Start) and Destination (End) roles: The verb "courir" in the utterance *couru de record en record/ran from record to record* in the reference was associated with the thematic grid End-Start-Theme. "courir" is a translation of "run" with the meaning of the *run-51.3.2* Verbnet class which has the role set Agent-Location-Theme. The arguments *from record* and *to record* do convey a Location sense, they were also correctly labeled with the End and Start roles but, based on the role sets it is using, our system can not detect this relation. Other roles which were confounded were Theme and Topic and Agent and Experiencer.

No appropriate Verbnet class. For 5 pairs there either was no appropriate Verbnet class or the Verbnet class did not have an appropriate thematic role. For instance there is no Verbnet class for "répondre" (answer) or "saisir" with the meaning

and the separation index gives information about the co-occurrence of objects and attributes (verbs and frames).

shown in *Les administrations sont saisies de propositions ....* Another example is the verb *dire*, which is assigned to the Verbnet class *say-37.7* by Sun *et al.* and is used with this meaning in the utterance *On peut dire autant de la confirmation .../One can say as much about the confirmation....* However, the *say-37.7* Verbnet class does not provide an appropriate role for *de la confirmation/about the confirmation.*

Syntactic frames missing in classification. In 5 cases the syntactic frame in the reference annotation was not associated with the verb by the classification and thus possibly prevented the association with the correct thematic grid.

Since the ⟨verb, thematic grid⟩ associations were produced by aligning groups of verbs with translated Verbnet classes, the quality of these associations is influenced by that of the translated Verbnet classes and by the mapping procedure. In the following we analyse the possible impact of these elements on the association of verb groups with thematic grids.

First, the groups of French verbs the FCA classes are aligned with are obtained by translating English Verbnet classes sharing a thematic role set. Looking at those classes which share a thematic role set, we found that one Verbnet thematic role set may be shared by many Verbnet classes: at least one and at most 21. Thus for example the Agent-End-Theme thematic grid is shared by 21 Verbnet classes comprising such semantically different classes as *butter-9.9* and *illustrate-25.3*. In contrast, 42 role sets (of 78) are specific to one Verbnet class with a relatively small number of verbs. This suggests that the approach of grouping Verbnet classes by thematic role sets may produce semantically inhomogeneous classes which may not properly reflect shared semantic components and is therefore less suitable for a semantic description of an FCA class[33]. It could therefore be more appropriate for creating the FCA class $\leftrightarrow$ thematic role set associations, to use the original translated Verbnet classes which are not grouped by thematic role sets. These would be semantically more coherent and their translations would better reflect their semantic meaning components, while still providing the link to the theta grids.

A further step in the association of FCA classes with thematic role sets is the alignment of the FCA classes with the translated Verbnet classes. Recall that each FCA class was associated with the thematic role set of the translated Verbnet class where the harmonic mean of the number of shared verbs and the size of the FCA class and the Verbnet class is highest[34]. A result of this approach was that often

[33]The distribution of verbs in the translated classes is similar in French and English.

[34]The FCA class C_{FCA} is associated with the thematic role set of the translated Verbnet class C_{VN} with highest F-measure between precision P and recall R, where $P = \frac{|C_{VN} \cap C_{FCA}|}{|C_{FCA}|}$, $R =$

(a) Syntactic arguments occuring in the associations produced based on the FCA classification.

AOBJ:PP	AOBJ:VPinf	
ATB:XP		
DEOBJ:PP	DEOBJ:VPinf	*DEOBJ:Ssub*
OBJ:NP		OBJ:Ssub
POBJ:PP		

(b) Syntactic arguments occuring in the reference annotations.

AOBJ:PP	AOBJ:VPinf	
ATB:XP		
DEOBJ:PP	DEOBJ:VPinf	
OBJ:NP	*OBJ:VPinf*	OBJ:Ssub
SUJ:NP		

Table 5.8: Syntactic arguments occuring in the associations produced based on the FCA classification (a) compared to syntactic arguments used in the reference annotations (b). In italics, the syntactic arguments which do not occur in both the reference and the FCA classification.

large inhomogeneous FCA classes were associated with small Verbnet classes with a very specific role set, since for these classes the proportion of shared verbs against the size of the classes is often best. Therefore a more sophisticated method mapping the FCA classes to translated Verbnet classes would also be desirable and could improve the associations with thematic role sets. A requirement to this procedure could be for example not to align classes which are too different in size or to associate to each FCA class at most one "best" thematic role set.

5.2.2.2 Syntax/semantics interface level.

In this section we analyse the associations of syntactic arguments with thematic roles induced by the FCA classification by using the SRL approach described in Section 5.2.1.2.

Of the semantic roles associated to the 427 ⟨verb, syntactic argument⟩ annotated types 399 (93.44%) were also associated with this pair by the FCA classification. For the 3605 annotated tokens, the semantic role associations of 3551 (96.28%) were also present in the FCA classification. That is, the FCA classification associated the same thematic role to a ⟨verb, syntactic argument⟩ pair as the manual annotation in a larger number of cases than the gold standard by [Sun *et al.*, 2010]: As pointed out in Section 5.2.2.1, 326 (76.34%) of the 427 ⟨verb, syntactic argument⟩ pairs in the reference were associated with a thematic role attributed to that verb by the gold by [Sun *et al.*, 2010].

Table 5.8a shows the syntactic arguments associated to verbs by this classification compared to the syntactic arguments annotated in the reference (Table 5.8b). The table shows that there are only few differences between the syntactic arguments associated to verbs by the FCA classifications and those annotated in the reference:

$\frac{|C_{VN} \cap C_{FCA}|}{|C_{VN}|}$.

SRL method	**%correct**			% of **incorrect**		%not labeled
	%total (Recall)	%labeled (Prec.)	F	%possible	%impossible	
baseline(default)	65.21	65.21	65.21	79.11	20.89	0
frame matching	28.35	68.00	40.02	83.16	16.84	58.31
SRL row 2 + prob.	34.17	48.29	40.03	88.10	11.90	29.24

Table 5.9: Results of the semantic role labeling approach, based on frame matching only and combined with probability estimates. The baseline is provided by default associations described in Section 5.2.1.2. Column "%not labeled" shows the proportion of instances which could not be labeled and columns "%possible" and "%impossible" give the proportion of incorrectly labeled instances which were ("%possible") or not ("%impossible") associated with the correct thematic role by the classification.

The DEOBJ:Ssub syntactic argument was produced from the FCA classification but was not among the annotated instances whereas OBJ:VPinf was among the annotated instances but is not associated to any verb by the FCA classification[35].

There were however 2085 (57.84%) ⟨verb, syntactic argument⟩ instances in the reference where the syntactic argument did not occur in any frame associated to the verb by the FCA classification and 286 distinct ⟨verb, syntactic argument⟩ pairs in the reference were not yielded by the FCA classification.

Table 5.9 gives an overview of the semantic role labeling results based on frame matching. Recall that the baseline is provided by a default assignment (as described in Section 5.2.1.2). The second row shows that the seed labeling produces an F-measure of 40.02, which is 25 points below the baseline. A high number of instances remained unlabeled (58.31%), but of the labeled instances 68% were correct, giving a precision of 3% above the baseline precision. In total the system labeled 1503 of 3605 (41.69%) instances and 1022 (67.99%) of these labels were correct. The columns captioned "%possible" and "%impossible" in Table 5.9 give details about the incorrectly labeled instances: They show the proportion of incorrectly labeled instances where the ⟨verb, thematic role⟩ association was present ("%possible") or not ("%impossible") in the classification. Thus, 83.16% of the incorrectly labeled instances were present in the classification (but not selected by the frame matching procedure). In these cases it would have been possible for an annotator to choose the correct role. As mentioned earlier, for 57.84% of the annotated ⟨verb, syntactic argument⟩ instances the syntactic argument did not occur in any frame associated to the verb by the FCA classification. Therefore the large number of instances which could not be labeled (58.31%) is mainly due to frames missing in the classification and less to the fact that the frame matching produced ambiguous associations.

For the second, probabilistic, step, we used the previously produced labels to

[35]There are 7 annotated instances with OBJ:VPinf syntactic arguments.

compute probability estimates P(semantic role|syntactic argument) and label those ⟨verb, syntactic argument⟩ instances which had not yet been assigned a thematic role with the semantic label with highest P(semantic role|syntactic argument). For these experiments we used a very small training corpus, namely the 3605 reference instances in the SRL gold corpus which were annotated by the frame matcher. We will further discuss the impact of the training size of the corpus later in this section. With this labeling, shown in row 3 of Table 5.9, about 1000 more instances could be associated with a role (2551 of 3605) but only about 210 of these labels were correct. The F-measure is very similar to that produced by frame matching only. Of the 1319 incorrectly labeled instances 1162 (88.10%) were also associated with the correct role. There was still a relatively large number of instances (29.24%) which could not be labeled. However, in both cases for a large proportion of incorrectly labeled instances (83.16% and 88.10%) the thematic role associated with the instance by the reference annotation was among the thematic roles mapped to the instance by the FCA classification.

Next, we compared these results to those of [Swier and Stevenson, 2005]. This is for information only, as the two approaches are similar but have different goals and use different data and therefore are not really comparable. Whereas Swier and Stevenson's goal was to produce an unsupervised SRL system using a manually constructed verb classification (Verbnet), in our work the aim was to use semantic role labeling to assess an automatically acquired verb classification. With respect to the data, the syntactic frame information used by Swier and Stevenson was obtained by parsing (or chunking) and is presumably less reliable than the syntactic frame information extracted from a treebank which we use in our method.

In the case of [Swier and Stevenson, 2005] the performance of the baseline lied roughly 9 points above the baseline's performance in our setting. Swier and Stevenson report a performance gain of 2 points using frame matching only, whereas in our setting the performance was 25 points below the baseline. The low performance was mainly due to the fact that a large proportion of ⟨verb, syntactic argument⟩ instances in the reference corpus were not generated by the classification and confirms what was already shown by the global evaluation in Section 5.2.2.1, namely that many (verb, frame) and (verb, thematic grid) pairs are missing from the classification. However, for those instances which could be labeled (42%), 68% were associated with the correct label, a precision which lies 3% above the baseline precision.

When adding probability estimates the F-measure of the labeling produced by our method stayed the same, whereas for Swier and Stevenson the increase is of 7 points. When using the annotations in the SRL gold corpus produced by the frame

[Swier and Stevenson, 2005]	SRL with FCA
subject	SUJ:NP
object	OBJ:NP OBJ:VPinf, OBJ:Ssub
indirect object	AOBJ:NP, AOBJ:VPinf, AOBJ:Ssub DEOBJ:NP, DEOBJ:VPinf, DEOBJ:Ssub
prepositional object	POBJ:PP, POBJ:VPinf, POBJ:Ssub
	ATB:XP (attribute)

Table 5.10: Syntactic arguments used in [Swier and Stevenson, 2005] and by our SRL method.

Experiment 1 complete	Experiment 2 partial	Experiment 3 basic
SUJ:NP	SUJ:NP	SUJ
OBJ:NP OBJ:VPinf OBJ:Ssub	OBJ:NP OBJ:VP	OBJ
AOBJ:NP AOBJ:VPinf AOBJ:Ssub	AOBJ:NP AOBJ:VP	AOBJ
DEOBJ:NP DEOBJ:VPinf DEOBJ:Ssub	DEOBJ:NP DEOBJ:VP	DEOBJ
ATB:XP	ATB:XP	ATB

Table 5.11: Syntactic arguments used in various experiments.

matcher, a larger number of instances could be labeled with our method but only one third of the additional labels were correct. [Swier and Stevenson, 2005] do not give the number of instances which could not be labeled by the frame matcher so we can make no comparison here.

There are however more differences between Swier and Stevenson's setting and ours. Thus the syntactic arguments in [Swier and Stevenson, 2005] are less specific than in our setting (Table 5.10).

Impact of syntactic argument representation. To explore the impact of the type of syntactic arguments used we repeated our experiments using two additional sets of less specific syntactic arguments. The first column in Table 5.11 shows the syntactic arguments used in the experiment described previously where the set of syntactic argument representation is the most specific. In the second experiment (column 2) we only distinguished between the syntactic categories NP (noun phrase) and VP (verbal phrase) and finally in the third experiment (column 3) we only took into account the grammatical functions. This latter type of syntactic arguments is closest to the one used by [Swier and Stevenson, 2005]. Table 5.12 shows the results

	Precision	Recall	F	Not labeled
Complete: most specific syntactic argument representations				
baseline	65.21	65.21	65.21	0
FAm	68.00	28.35	40.02	58.31
FAm + prob	48.29	34.17	40.03	29.24
Partial: less specific syntactic argument representations				
baseline	65.21	65.21	65.21	0
FAm	70.40	30.87	42.92	56.14
FAm + prob	63.99	47.32	54.41	26.05
Basic: basic syntactic argument representations				
baseline	65.21	65.21	65.21	0
FAm	77.84	28.16	41.35	63.83
FAm + prob	69.87	30.04	42.02	57.00

Table 5.12: SRL results when using various types of syntactic arguments. *Fam* represents the labeling obtained when using frame matching only and *FAm+prob* by using probability estimates computed using the results of frame matching on the SRL gold instances. Labeling results are compared to the reference annotations (SRL gold).

compared to the reference for the SRL obtained using these three different sets of syntactic arguments.

The F-measures when using these less detailed syntactic argument representations were similar to that of the system with the most specific syntactic argument representation. The number of instances which could not be labeled was slightly lower for the less specific argument representation and much higher for the basic syntactic argument representation. On the other hand, precision was substantially higher for the basic syntactic argument representation and slightly higher for the less detailed syntactic argument representation, mirroring the number of unlabeled instances.

In general the number of instances which could not be labeled by the frame matcher was relatively large: 56.14% – 63.83%. Most of these instances (2085 of 2102 for the most detailed syntactic argument representation and 2085 of 2301 for the simplest representation corresponding to 286 respectively 262 distinct ⟨verb, syntactic argument⟩ pairs) could not be labeled because the ⟨verb, syntactic argument⟩ pairs could not be derived from the FCA classification: none of the frames associated to the verbs contained the syntactic argument. The reason for this may be that either no frame containing this syntactic argument was associated to the verb by the lexicon, or that the class giving this association was discarded in the filtering process. Using a more detailed representation of syntactic arguments the following arguments could be labeled: AOBJ:PP, OBJ:NP and SUJ:NP. For these arguments we can compute probability estimates which we can use in the next labeling step. In

contrast, when using the simplest syntactic argument representation the syntactic arguments which could be labeled are SUJ and AOBJ.

When combining information from lexical data and probability estimates, using a more detailed syntactic argument representation decreased the number of instances which could not be labeled by almost 30%, but for the less specific representation the decrease was of only about 7%. For a possible explanation consider the following: The number of instances which can be labeled with the help of probability estimates depends on the distinct syntactic arguments, occuring in the corpus which could be labeled in the previous frame matching step (which in turn did not depend on the corpus, but on the FCA classification built from the lexicon). As we saw earlier, using the "complete" syntactic argument representation and based on the labeling obtained from the FCA classification we could compute probability estimates for the following syntactic arguments: AOBJ:PP, OBJ:NP and SUJ:NP. As these are frequent syntactic arguments, we may expect a large number of instances to be labeled. In contrast, using the "basic" syntactic argument representations we only could compute probability estimates for AOBJ and SUJ, so the combined system will for example not be able to disambiguate instances containing the OBJ syntactic argument (which is very frequent).

Surprisingly, in the case of the "partial" syntactic argument representation the F-measure of semantic role labeling based on the combination of the FCA data and the probability estimates increased by approximately 10 points. The reason for this is the following. The same syntactic arguments as with the more detailed syntactic argument representation were labeled in the frame matching step, namely AOBJ:PP, OBJ:NP and SUJ:NP. Most OBJ:NP instances (113) were associated with the Theme role, whereas in the case of the more detailed syntactic argument representation most OBJ:NP instances (124) were associated with the Patient role – which was obviously correct in fewer cases compared to the reference. Therefore, in this case, the increase in F-measure must be considered an idiosyncrasy of the data set under consideration.

Overall, the probability estimates did help to improve the F-measure of the semantic role labeling, but the achieved F-measures were below the baseline. A rather large number of additional instances could be labeled, but a large proportion of these labels were incorrect. So, while the classification allowed to label roughly 40% of the reference corpus instances and about 60% of these assignments were correct, these associations did not provide sufficient information to correctly predict the labels for the remaining unlabeled instances. However, the idiosyncrasies observed in the previous discussion suggest that the system using the probability estimates is suffering from lack of data.

	Precision	Recall	F	Not labeled
Complete: most specific syntactic argument representations				
baseline	65.21	65.21	65.21	0
FAm-1	70.24	25.99	37.94	63.00
FAm-1 + prob	59.77	57.23	58.46	4.30
Partial: less specific syntactic argument representations				
baseline	65.21	65.21	65.21	0
FAm-1	69.61	25.99	37.85	62.66
FAm-1 + prob	59.04	54.09	56.46	8.38
Basic: basic syntactic argument representations				
baseline	65.21	65.21	65.21	0
FAm-1	69.81	25.91	37.79	62.89
FAm-1 + prob	61.30	53.04	56.87	13.48

Table 5.13: SRL results when using various types of syntactic arguments and with variant 1 of frame matching method. *Fam-1* rows show labeling results when using frame matching only and *Fam-1+prob* show the labeling obtained when computing probability estimates based on the frame matching results on the SRL gold instances. Labeling results are compared to the reference annotations (SRL gold).

Impact of frame matching method. In this first series of experiments we used the more restrictive frame matching method described in Section 5.2.1.2. In the next section we explore the impact of the frame matching method on the two labelings, one based on the FCA classification only, the other on a combination of the FCA classification and corpus data, by applying the less restrictive frame matching method presented in Section 5.2.1.2. This variation made use of the FCA classes and matched the subcategorisation frames in each FCA class with the thematic grids corresponding to this FCA class. Since in the method we used up to now all subcategorisation frames associated with a verb were matched with all thematic grids associated to this same verb, this second method is less restrictive.

We show the results in terms of precision, recall, F-measure and percentage of instances not labeled for this variant of the frame matching method (*FAm-1*) in Table 5.13. We observe that the F-measure achieved by the SRL based on this less restrictive frame matching only is lower than when using the more restrictive frame matching method (2-4 points). The number of instances which could not be labeled in the frame matching step is more important than for the more restrictive frame matching when using the more specific syntactic argument representations (approximately 5 points) and is similar in the case of the less specific syntactic argument representation. When combining the frame matching with probability estimates the produced labeling has an F-measure of at most 58.46 vs. 54.41 for the more restrictive method. In contrast, we do not observe the same differences for the different representations of syntactic arguments as when applying the more

restrictive frame matching method: here the best F-measure and the most instances labeled were obtained for the most specific syntactic argument representation and the performance was comparable for each type of syntactic argument representation. This is due to the fact that for each syntactic argument representation a sufficient number of distinct syntactic arguments could be labeled:

complete: AOBJ:PP, DEOBJ:PP, OBJ:NP, OBJ:Ssub, POBJ:PP, SUJ:NP

partial: AOBJ:PP, DEOBJ:PP, OBJ:NP, OBJ:VP, POBJ:PP, SUJ:NP

basic: AOBJ, DEOBJ, OBJ, POBJ, SUJ

To sum up, this suggests that with the more restrictive method we obtain a more accurate labeling but too few distinct syntactic argument instances could be labeled, such that the counts for these labeled syntactic arguments did not help at predicting roles for the not yet labeled instances. This behaviour of the labeler was more pronounced in the case where the syntactic argument representation is less detailed, ie. when there are less distinct syntactic arguments to be labeled. In this case there were more ambiguities (because more thematic roles are assigned to less syntactic arguments) and less instances could be labeled in the frame matching step. With the less restrictive frame matching method, precision increased and recall decreased. This was probably caused by the frames missed by the classification, but it also suggests that many of the associations of verbs with subcategorisation frames given by the FCA concepts did not match those occurring in the reference. Although with this method the labeling produced by frame matching only was less accurate, there was an important improvement for all types of syntactic argument representations when using the probability estimates. Therefore this noisier frame matching allowed a larger gain in F-measure for the system combining lexical and corpus data, but the F-measure as well as the precision (proportion of correct in labeled instances) stayed well below the baseline.

In all these experiments we only labeled the annotated corpus instances and then used the labeled instances for training. The results, in particular those obtained with the more restrictive frame matching method, showed that relatively few distinct syntactic arguments could be labeled, thus hampering the next labeling step. Therefore these results, more specifically those of the labeling using probability estimates, might be improved by using more corpus data – hopefully thus more syntactic arguments may be labeled in the frame matching step allowing to produce probability estimates for more syntactic arguments.

(a) Using a detailed syntactic argument representation.

		restrictive frame matching			less restrictive frame matching		
	BL	FAm	FAm+prob	FAm+p7+prob	FAm1	FAm1+prob	FAm1+p7+prob
F	65.21	40.02	40.03	38.89	37.94	58.48	58.35
not labeled	0	58.31	29.38	22.69	63.00	4.30	3.86

(b) Using a less detailed syntactic argument representation.

		restrictive frame matching			less restrictive frame matching		
	BL	FAm	FAm+prob	FAm+p7+prob	FAm1	FAm1+prob	FAm1+p7+prob
F	65.21	42.92	54.41	42.22	37.85	56.46	56.46
not labeled	0	56.14	26.05	21.03	62.66	8.38	8.38

(c) Using a basic syntactic argument representation.

		restrictive frame matching			less restrictive frame matching		
	BL	FAm	FAm+prob	FAm+p7+prob	FAm1	FAm1+prob	FAm1+p7+prob
F	65.21	41.35	42.02	38.89	37.79	56.87	57.50
not labeled	0	63.83	57.00	22.69	62.89	13.48	16.09

Table 5.14: Results when performing the labeling for all ⟨verb, syntactic argument⟩ pairs in the P7 corpus: When using the more (FAm) or less (FAm1) restrictive frame matching method only or combined with corpus data. Corpus data consists of either the SRL gold instances (FAm/FAm1+prob) or all P7 corpus instances (FAm/FAm1+p7+prob). The ⟨verb, syntactic argument, thematic role⟩ associations are compared to those in SRL gold.

Impact of training corpus size. We explored this direction of improvement by producing semantic role labelings of two corpora: The first corpus consisted of the ⟨verb, syntactic argument⟩ instances of the annotated P7 instances only (the SRL gold instances, as before) whereas the second corpus consisted of all the ⟨verb, syntactic argument⟩ instances of the P7 corpus (see Section 3.2.1.2 for a brief description of the SRL gold and P7 corpus). The assumption we wanted to verify was, that using all P7 instances there is no important decrease in F-measure, but more instances can be labeled. If this assumption is verified, using all P7 instances would be useful in this semantic role labeling task.

Table 5.14 shows the results of these experiments. We used the two frame matching methods discussed above: First the more restricted method where a subcategorisation frame associated to the verb is mapped with all the thematic grids associated to that verb (FAm), and second the less restrictive method where the frames associated to a verb by an FCA class are mapped to the thematic grids associated to that verb by the FCA classification (FAm-1). As before we used three syntactic argument representations, where the most detailed one is the syntactic argument representation also used in the P7 corpus. For the most detailed syntactic argument representation (Table 5.14a) our assumption was confirmed: When labeling the entire P7 corpus and using the resulting counts to compute probability estimates F-measure decreased slightly and the number of labeled instances increased. However, this increase was

much more important for the more restrictive frame matching method (roughly 5% vs. only about 0.5%).

The situation was different for the experiments with the less detailed syntactic argument representation: For the more restrictive frame matching method, the assumption was supported by an increase of labeled instances of about 5%. However, there was an important decrease in F-measure (approximately 10), such that the F-measure when using P7 corpus data was slightly lower than the F-measure obtained by using frame matching only. Looking at the data we observed the following: First note that as we saw earlier the more restrictive frame matcher associated most ⟨verb, OBJ:NP⟩ pairs with the Patient role and this association corresponded to the annotated corpus instances. By labeling all of the P7 instances with the more restrictive frame matcher, as assumed, more syntactic arguments could be labeled. However, in most cases the labeler associated ⟨verb, OBJ:NP⟩ pairs with the Theme role. This does not correspond to the OBJ:NP usage in the annotated instances, hence the important decrease in F-measure. This observation highlights the role of the corpus in this semantic role labeling method: Based on the FCA classification, independently from the corpus, the OBJ:NP syntactic argument could only be associated unambiguously with the Theme role. When computing probability estimates from the annotated instances as corpus, OBJ:NP was associated with the Patient role. This shows, that using the probability estimates it was possible to correct the prediction produced based on the FCA classification only. In contrast, when labeling the entire P7 corpus and computing probability estimates from this data, OBJ:NP was associated with the Theme role, which engendered the lower F-measure (compared to when using the annotated instances as corpus). It also shows that in this respect the annotated instances are not representative of all of P7. In addition there are more factors which may have played a role in this outcome:

- In the case of the annotated instances, syntactic arguments were corrected manually and are therefore more reliable than in the rest of P7.
- The Theme and Patient roles are particularly difficult to distinguish, therefore the annotations may not be always accurate.
- The OBJ:NP syntactic argument may be a realisation of a particularly large number of thematic roles. For example, in our annotations OBJ:NP was labeled with the following 9 roles: Agent, End, Experiencer, Location, Patient, PredAtt, Start, Theme, Topic.

In contrast, when using the less restrictive frame matching method results were the same for both corpora.

Regarding the less detailed syntactic argument representation, the results confirmed again our assumption for the more restrictive frame matching method: When using the P7 corpus, there was a small decrease in F-measure (of about 3 points) but an important increase in the percentage of labeled instances (approximately 24%). The results did not seem to exhibit the same pattern for the less restrictive frame matching method: When using P7, there was a slight increase in F-measure compared to using the SRL gold corpus but the number of labeled instances decreased by roughly 3%. Looking at the data we found that in both cases, using the annotated data and the entire P7 as corpus, the same syntactic arguments were labeled: SUJ, OBJ, AOBJ, DEOBJ and POBJ. However, on the P7 corpus POBJ was labeled the same number of times with the roles Theme and Agent, so in this case the probability estimates could not be used to associate POBJ with an unambiguous thematic role. Therefore this deviation from the pattern was due to an idiosyncrasy of this corpus[36].

Overall these results confirmed our assumption that the P7 data may help to label more instances. They also show that ultimately its effects also depend on several other factors, as for example the frame matching method, the syntactic argument representation scheme and the corpus. For example, using P7 data had only little effect on the results when using the less restrictive frame matching method, in particular it had no effect for the less detailed syntactic argument representation. Therefore it should maybe best be determined empirically with respect to the data at hand and the specific needs of the application.

We show a more detailed analysis of the performance of our method (based on the FCA classification) with respect to involved semantic roles, syntactic arguments, number of subcategorisation frames and polysemy classes in Appendix B.

5.2.3 The IGNGF Clustering

In this section we present results of the evaluation of IGNGF clusterings. As for the FCA based classification we evaluated the classification produced in two ways. First we evaluated the verb, syntactic frame and verb, thematic grid associations it provides. Second, we assessed its ability to support the correct assignment of thematic roles to verb arguments.

As described in Section 3.3.2, the IGNGF clustering method produced a (non-overlapping) grouping of the French verbs in data set **vn_all**, based on the feature set **grid-scf-sem**, consisting of subcategorisation frames, additional semantic features

[36] However this kind of tie may occur easily, considering the relatively small number of distinct syntactic arguments we are dealing with.

and features derived from the translated Verbnet classes. To associate verb clusters both with a set of subcategorisation frames and a set of thematic roles (Verbnet classes), we explored the following four approaches:

Fmax We associate to each cluster the cluster maximising subcategorisation frames (features).

Fpos We associate to each cluster the features (subcategorisation frames) with Feature F-measure above a global threshold (see Section 3.3.2.4).

theta-1 We associate to each cluster the thematic role sets which occur in its feature set.

theta-2 We associate to each cluster the thematic role set(s) of the translated Verbnet class(es) maximising the f-measure of the verbs they have in common[37].

To determine the "best" way of associating the verb clusters with syntactic frames and thematic grids, we built classifications for each of the configurations above and evaluated them by checking them on the SRL gold corpus as described in Section 5.2.1.2. Thus, the process of finding the best configuration for the clustering was in effect guided by the assessment of the syntax-semantic interface induced by the classification. Therefore we start the IGNGF clustering evaluation with this issue.

5.2.3.1 Syntax/semantics interface level.

We first determined the best way to associate verb clusters with thematic grids as follows. As for the evaluation in Section 4.2.2, verb clusters were associated with the subcategorisation frame features which maximise feature F-measure (**Fmax** association scheme). Verb clusters were then associated with thematic grids in the two ways listed above. We then evaluated the resulting classification on the SRL gold corpus by assessing the assignments of thematic roles to ⟨verb, syntactic argument⟩ in SRL gold induced by the frame matching based on the IGNGF classification. As the results in Table 5.15 show, the classification obtained by assigning the thematic grids to the clusters using the translated Verbnet classes rather than the cluster features produced better ⟨verb, syntactic arguments, thematic role⟩ associations (better F-measure and less unlabeled instances). Therefore, for the following experiments

[37] If C_{VN} is a translated Verbnet class and C_{IGNGF} the set of verbs in an IGNGF cluster, recall $= \frac{|C_{\text{VN}} \cap C_{\text{IGNGF}}|}{|C_{\text{VN}}|}$, precision $= \frac{|C_{\text{VN}} \cap C_{\text{IGNGF}}|}{|C_{\text{IGNGF}}|}$ and f-measure is the harmonic mean of precision and recall: $\frac{2 * \text{precision} * \text{recall}}{\text{precision} + \text{recall}}$.

SRL method	**correct**			% of **incorrect**		%not labeled
	%total (Recall)	%labeled (Prec.)	F	possible	impossible	
theta-1	13.01	79.36	22.35	0	100	83.61
theta-2	30.26	76.72	43.41	12.39	87.61	60.55

Table 5.15: SRL results when associating verb clusters with thematic grids by using the cluster features (row *theta-1*) and the translated Verbnet classes (row *theta-2*). The association with subcategorisation frames is obtained using cluster maximising features and the assigned thematic roles are those occuring among the cluster features (*theta-1*) or obtained by aligning the clusters with the translated Verbnet classes (*theta-2*). The labeling with thematic roles is produced using frame matching only.

SRL method	**correct**			% of **incorrect**		%not labeled
	%total (Recall)	%labeled (Prec.)	F	possible	impossible	
Fmax	30.26	76.72	43.41	12.39	87.61	60.55
Fpos	47.02	72.53	57.05	18.07	81.93	35.17

Table 5.16: SRL (frame matching) results when using the IGNGF clustering. The association with subcategorisation frames is obtained using features with f-measure above a global threshold and the thematic roles are assigned by aligning the clusters with the translated Verbnet classes. Row Fmax shows results when associated SCFs are maximising cluster Feature F-measure and row Fpos when associated SCFs are those whose F-measure either is above a global threshold for the overall clustering.

we associated the verb clusters with thematic grids this way. Yet, this configuration resulted in a large number of instances which could not be labeled (60.55%). Looking at the syntactic argument types which could be labeled in the frame matching step we found that only AOBJ:PP, OBJ:NP and SUJ:NP could be assigned unambiguous thematic roles. We suspected that better results could be achieved if a larger set of distinct syntactic arguments could be labeled in the frame matching step. To explore this direction, in the next series of experiments we assigned subcategorisation frames to clusters following the **Fpos** cluster labeling scheme: in addition to the cluster maximising frames, clusters were associated (labeled) with the features with a Feature F-measure above a global threshold, which is the average of the Feature F-measure of all the cluster maximising features (see Section 3.3.2.4). Table 5.16 shows the result for this series of experiments: The clustering was the same as before but a cluster was associated with all the SCFs where the feature F-measure was above a global threshold[38] Thematic role associations were obtained by alignment with the translated Verbnet classes as before. In this configuration the results improved considerably: The f-measure for the labeling built from the clustering only was around 57 points and the number of instances which could not be labeled was

[38]In most cases cluster maximising features have an F-mesure above this threshold. This F-measure may however be below this threshold because of the normalisation scheme we applied.

(a) Complete: most detailed syntactic argument representation

SRL method	**correct**			% of **incorrect**		%not labeled
	%total (Recall)	%labeled (Prec.)	F	possible	impossible	
baseline	65.21	65.21	65.21	79.11	20.89	0.00
FM	47.02	72.53	57.05	18.07	81.93	35.17
prob	61.72	63.96	62.82	36.52	63.48	3.50
P7prob	61.72	63.94	62.81	36.49	63.51	3.47

(b) Partial: less detailed syntactic argument representation

SRL method	**correct**			% of **incorrect**		%not labeled
	%total (Recall)	%labeled (Prec.)	F	possible	impossible	
baseline	65.21	65.21	65.21	79.51	20.49	0.00
FM	47.41	72.69	57.39	18.07	81.93	34.79
prob	62.39	64.07	63.22	29.66	70.34	2.64
P7prob	62.39	64.06	63.21	29.64	70.36	2.61

(c) Basic: basic syntactic argument representation (syntactic functions only)

SRL method	**correct**			% of **incorrect**		%not labeled
	%total (Recall)	%labeled (Prec.)	F	possible	impossible	
baseline	65.21	65.21	65.21	79.98	20.02	0.00
FM	47.43	71.91	57.16	0.00	100.00	34.04
prob	62.80	64.45	63.61	0.00	100.00	2.55
P7prob	62.58	64.22	63.39	0.00	100.00	2.55

Table 5.17: SRL results when using the IGNGF clustering, more detailed (a), less detailed (b) or basic (c) syntactic argument representation. The *FM* rows show the results for the labeling obtained using frame matching only, *prob*, using frame matching and probability estimates and for *P7prob* probability estimates are computed from the P7 corpus.

also lower than in the other configurations. We therefore used this configuration for the following experiments.

Finally we conducted the same experiments for IGNGF as for FCA, as presented in Section 5.2.2.2: We used frame matching and the IGNGF clustering to assign unambiguous thematic roles to ⟨verb, syntactic arguments⟩ pairs. Then, we computed probability estimates based on the produced labeling using the annotated P7 instances (*prob*) or all P7 instances (*P7prob*). As for the FCA classification (Section 5.2.2) we performed each experiment with three more or less detailed syntactic argument representations. Since the IGNGF clusterings are crisp (non overlapping) the less restrictive frame matching method performs the same way as the more restrictive one and was therefore not used in these experiments. The results are shown in Table 5.17. The table shows that the performance of the systems was similar, regardless of the type of syntactic argument representation used. The F-measure for the FM only SRL system was around 57 points, with around one third of corpus instances which could not be assigned a label. This performance was below the default

(a) Distribution of ⟨verb, frame⟩ pairs in various resources. SCFs are associated using the *Fpos* labeling scheme alone or with an additional special treatment for frequent frames (*Fpos+trans*).

SCFs	SRL gold	classif	SRL gold & classif	SRL gold & lex ¬ classif	SRL gold ¬ lex	Recall
Fpos	316	1100	134	140	42	42.41
Fpos+trans	316	1149	163	111	42	51.58

(b) Distribution of ⟨verb, grid⟩ pairs in various resources.

Grids	SRL gold	SRL gold & classif	Recall
types	318	124	38.99
tokens	1600	731	45.69

Table 5.18: Distribution of ⟨verb, frame⟩ pairs 5.18a and ⟨verb, grid⟩ pairs 5.18b in various resources.

baseline but the precision was high, around 72%, for each type of syntactic argument representation. As for FCA, F-measure could be increased (to about 63) and almost all instances could be labeled when using corpus data. However, when using corpus data, precision decreased about 2% below the default baseline precision. Yet, the fact that almost all instances could be labeled suggests that the frame matching could assign roles to almost all syntactic arguments occuring in the reference corpus instances. The gain when using the entire P7 corpus data was negligible.

The three systems perhaps showed only one major difference: for the simplest syntactic argument representation (Table 5.17c) and the incorrectly labeled instances, the correct roles were not associated with these instances at all, whereas for the more detailed syntactic argument representations (Table 5.17a and 5.17b) about 1/3 of these incorrect instances were associated with the correct roles via the clustering (*possible* vs. *impossible* columns in the table).

5.2.3.2 Global level evaluation.

As before for the FCA classification, for the global level evaluation, we compared the frames and grids present in the SRL gold corpus with those given by the classification. Table 5.18a (*Fpos* row) shows for the ⟨verb, syntactic frame⟩ pairs present in the reference, the proportion of pairs also present in the classification and in the syntactic lexicon we started from. The recall for frames was 42.41% for types (37.62% for tokens). As the table shows, there wer many more frames in the classification than in the SRL gold corpus hence we do not compute precision. When omitting the 42 ⟨verb, frame⟩ pairs in the reference, which were not present in the initial lexicon, recall increased to 48% (134/274). There were however 140 (verb,frame) pairs present in the SRL gold corpus which were not in the clustering (but in the lexicon). Investigating

reasons for this we found that very common frames such as in particular the transitive frames, had a low feature recall (they failed to clearly distinguish a cluster from other clusters) so that they might not have been included in the set of features characterising a cluster. This occurs in particular for clusters which are not very homogeneous and where these common frames have both a low feature recall and a low feature precision (they are not shared by many verbs in the cluster). Thus out of the 134 missing pairs, 57 included the transitive frames (SUBJ:NP;OBJ:NP) and 18 the transitive frame with a prepositional object. By consequence, we adjusted the cluster labeling process to account for these cases by associating the clusters with frames which are shared by more than 70% of the verbs in the cluster. This resulted in the scores shown in Table 5.18a, row *Fpos+trans*.

We see that recall increased considerably. However, of 111 missing pairs, 29 still included the transitive frame and 26 the transitive frame with a prepositional object, suggesting that the cluster labeling process needs to be further refined to account for these very frequent frames.

Other defective cases appeared to be cases where two Verbnet classes had been grouped into one so that the syntactic frames labeling this "double class" were in fact characteristic for almost disjoint subsets, but were still wrongly assigned to all the verbs. An example of such a cluster is a cluster which groups together verbs from the *amuse-31.1* Verbnet class (Cause-Experiencer) and from the *modes_of_being_with_motion-47.3* class (Agent-Location-Theme). For this cluster, the most relevant syntactic frames are SUJ:Ssub,OBJ:NP (*Qu'il parte ennuie Marie/His leaving bothers Marie*) and SUJ:NP,DEOBJ:NP (*Il tremble de froid/He is trembling from cold*) which both have relatively low feature F-measure. Of these frames the first is only valid for *amuse-31.1* verbs and the second only for *modes_of_being_with_motion-47.3* verbs.

Finally, since we used a hard (non overlapping) clustering method, verbs only were assigned to one class. As a result, polysemous verbs often failed to be assigned the distinct frames and grids associated with their distinct senses. For example the verb "briser" (*break, wreck*), which has both an Agent-Instrument-Patient (*Le choc a brisé la statue/The shock broke the statue.*) and a Cause-Experiencer (*Le départ de Max a brisé Marie/Max's leaving broke/wrecked Mary.*) sense, was a member of a cluster labeled with the Agent-Instrument-Patient role set and with syntactic frames valid for both meanings. Although the frames which are most distinctive for the Cause-Experiencer meaning[39] had lowest feature F-measure, our cluster labeling techniques were currently too coarse grained to make these sense distinctions explicit.

[39]These are the frames where the subject may be sentential: SUJ:Ssub,OBJ:NP and SUJ:Ssub,POBJ:PP.

Considering now the associations with thematic grids, Table 5.18b shows the distribution of ⟨verb, grid⟩ pairs in the various resources. The column *SRL gold & classif* lists the number of ⟨verb, grid⟩ pairs (types) in the corpus, for which the verb was associated with a "compatible" grid by the classification, that is the roles of the classification grid were a super-set of the corpus grid roles. According to this, only 124 of the 318 ⟨verb, grid⟩ pairs present in the reference were present in the classification. One reason for this discrepancy is polysemous verbs. As mentioned above, roughly 21% of the verb instances in the corpus were used with a meaning different from that captured by the 12 Verbnet classes from which the sample verbs were taken. Since IGNGF only allows for a given verb to be in one cluster (hard, non overlapping clustering), the possible verb/thematic grid associations are restricted.

5.2.4 IGNGF vs. FCA

In this section we compare the IGNGF and FCA classifications with respect to the evaluation of ⟨verb, frame⟩ and ⟨verb, grid⟩ associations described in Section 5.2.2 and 5.2.3. For both types of classification methods we used the most detailed syntactic argument representation and the configuration performing best. The discussion of the FCA vs. the IGNGF classification follows the same schema as for each of the classifications. We first compare the classifications on the global level and then discuss their performance on the syntax/semantics interface level. We found that overall the FCA classification performed better with respect to the associations of verbs with frames and grids (global evaluation) whereas the IGNGF clustering better supported the assignment of thematic roles to syntactic arguments (evaluation on the syntax/semantics interface level). In the following we further detail the comparison on both levels.

5.2.4.1 Global level evaluation.

Table 5.19a shows a comparison of frames present in the SRL gold corpus and those produced by the two classifications.

The table shows that the recall was substantially higher for the FCA classification, that is the FCA classification contained a larger number of reference ⟨verb, frame⟩ associations than the IGNGF classification. Along the same lines, the number of reference ⟨verb, frame⟩ pairs, present in the lexicon but missed by the classifications was substantially larger for IGNGF than for FCA (111 of 316 for IGNGF, vs. 31 of 316 for FCA). The FCA classification produced a much larger number of ⟨verb, frame⟩ pairs than the IGNGF classification, but as we can not evaluate precision

(a) Distribution of ⟨verb, frame⟩ pairs in various resources, for the IGNGF and FCA classifications

SCFs (types)	SRL gold	classif	SRL gold & classif	SRL gold & lex ¬ classif	SRL gold ¬ lex	Recall	Recall w/o missing in lex
IGNGF	316	1149	163	111	42	51.58	59.59
FCA	316	16542	243	31	42	76.90	88.69

(b) Number of correct ⟨verb, frame⟩ pairs in FCA and IGNGF classifications (and both).

SRL gold ∩ lex	IGNGF ∩ SRL gold	FCA ∩ SRL gold	FCA ∩ IGNGF ∩ SRL gold
274	163	243	147

Table 5.19: Distribution of ⟨verb, frame⟩ pairs in gold and FCA and IGNGF classifications: separately (a) and simultaneously (b).

aller	SUJ:NP,AOBJ:PP
aller	SUJ:NP,OBJ:VPinf
associer	SUJ:NP,OBJ:NP,AOBJ:PP,POBJ:PP
concourir	SUJ:NP,AOBJ:PP
contribuer	SUJ:NP,AOBJ:VPinf
dire	SUJ:NP,OBJ:NP,ATB:XP
définir	SUJ:NP,OBJ:NP,ATB:XP
garder	SUJ:NP,OBJ:NP,ATB:XP
heurter	SUJ:NP,AOBJ:PP
incorporer	SUJ:NP,OBJ:NP,AOBJ:PP
participer	SUJ:NP,AOBJ:PP
passer	SUJ:NP,AOBJ:PP,DEOBJ:PP
porter	SUJ:NP,AOBJ:PP
voir	SUJ:NP,ATB:XP
voir	SUJ:NP,OBJ:Ssub,ATB:XP

Table 5.20: The 15 reference ⟨verb, frame⟩ pairs missing in both the FCA and IGNGF classification.

(because the reference associations are not exhaustive) we can not estimate which is more correct. Table 5.20 shows the 15 reference ⟨verb, frame⟩ pairs which were missing in both the FCA and the IGNGF classification. We see that many subcategorisation frames in these pairs were missing in the FCA classification, namely those containing AOBJ:PP and ATB:XP arguments (see Section 5.2.2.1). Interestingly, as Table 5.20 shows, these were also often lost by the IGNGF classification, so both classification methods had problems to accurately represent the information given by the associations with these frames provided by the lexicon. For both classifications, associations of verb classes with these subcategorisation frames were not coherent with respect to the classification criteria of the respective classification method. Possibly the lexical information for these frames needs to be enhanced by additional supporting features (semantic features for AOBJ:PP or ATB:XP).

Next we comment on the ⟨verb, grid⟩ associations. Table 5.21a shows, for the IGNGF and FCA classification the number of corpus ⟨verb, grid⟩ instances for which

(a) Distribution of ⟨verb, grid⟩ pairs in various resources, for the IGNGF and FCA classifications.

Grids	gold	gold & classif	R
IGNGF	318	153	48.11
FCA	318	280	88.05

(b) Number of reference ⟨verb, grid⟩ pairs compatible with the FCA, IGNGF or both classifications.

gold	IGNGF ∩ gold	FCA ∩ gold	FCA ∩ IGNGF ∩ gold
318	153	203	149

Table 5.21: Distribution of ⟨verb, grid⟩ pairs in various resources: for the IGNGF and FCA classifications separately (a) and simultaneously (b).

the grids were compatible with the grids associated to the verb by the classifications. That is, for these grids, the role set present in the reference was a subset of a role set associated with the verb by the classification. As the table shows, the recall for the FCA classification was much higher than for the IGNGF classification, that is, many more reference ⟨verb, grid⟩ pairs are compatible with ⟨verb, grid⟩ associations produced by the FCA classification than pairs generated by IGNGF. Table 5.22 shows the 34 (of 318) reference ⟨verb, grid⟩ pairs which were compatible with neither the FCA nor the IGNGF classification (ie. the grid was not compatible with any grid associated with the verb by any of the classifications). Most of the reference ⟨verb, grid⟩ pairs which were incompatible with FCA, were also incompatible with the IGNGF classification (34 of 38). In Section 5.2.2.1 we discussed possible reasons for these missing associations: These were related to annotation issues, semantic role and/or Verbnet class confusions and also possibly to the missing associations with syntactic frames. Table 5.21b shows the proportion of reference ⟨verb, grid⟩ pairs produced by both the FCA and IGNGF classification: Of the 153 correct pairs in the IGNGF classification 149 (97.36%) were also present in the FCA classification, which supports the coherence of these verb group and thematic grid associations.

5.2.4.2 Syntax/semantics interface level.

We now compare the IGNGF clustering and the FCA classification with respect to their ability to associate ⟨verb, syntactic argument⟩ instances in the reference corpus with thematic roles. As described in Section 5.2.1.2, we used each classification to assign theta-roles to the ⟨verb, syntactic argument⟩ pairs in the SRL gold corpus and then compare these associations with those in the reference (SRL gold). This semantic role assignment was produced first based on the lexical resources only using

affirmer	Cause-Theme
aller	End-Start-Theme
aller	End-Theme
appréhender	Agent-Topic
conserver	Agent-Beneficiary-Theme
contribuer	Agent-Theme
contribuer	Agent-ThemeSym
contribuer	Agent-ThemeSym-ThemeSym
contribuer	ThemeSym-ThemeSym
corriger	Instrument-Theme
courir	End-Start-Theme
dire	Agent-Theme-Topic
désigner	Agent-PredAtt-Theme
désigner	Cause-Topic
désigner	PredAtt-Theme
empiler	Location-Theme
examiner	Experiencer-Theme
écouter	Experiencer-Theme
fortifier	Agent-Instrument-Patient
incorporer	Agent-PatientSym-PatientSym
inonder	Agent-Instrument-Theme
inonder	Instrument-Theme
lancer	Instrument-Theme
osciller	End-Start-Theme
participer	End-Theme
prendre	Agent-PredAtt-Theme
préserver	Instrument-Theme
ressentir	Experiencer-Theme
répondre	Theme-Topic
réserver	Agent-Beneficiary-Theme
saisir	Agent-Theme-Topic
saisir	Experiencer-Topic
scruter	Experiencer-Theme
supprimer	Cause-Theme

Table 5.22: The 34 reference ⟨verb, grid⟩ pairs which are not compatible with any of the FCA and IGNGF classifications.

the frame matching method and second on a combination of frame matching and probability estimates (see Section 5.2.1.2). Results are summarised in Table 5.23. Table 5.23a shows a comparison of FCA and IGNGF based on frame matching only and in Table 5.23b we present the results for the combined labeling. We see that overall the IGNGF clustering outperformed the FCA classification in terms of F-measure and number of labeled instances. With respect to the labeling based on frame matching only, we obtained best results (in terms of F-measure) for both classifications when using a less detailed syntactic argument representation. With an F-measure of 56.14% vs. 42.92% and less unlabeled instances (34.79% vs. 56.15%) the method based on the IGNGF clustering clearly outperformed the one based on FCA. For both systems, the F-measure is below the baseline. However, for both

(a) Results for ⟨verb, syntactic argument, theta role⟩ associations based on IGNGF cluster vs. FCA classification when using frame matching only.

	correct			% of **incorrect**		%no label
	%total (R)	%labeled (P)	F	possible	impossible	
baseline	65.21	65.21	65.21	79.11	20.89	0.00
FCA (partial)	30.87	70.40	42.92	82.05	17.95	56.14
IGNGF (partial)	47.43	71.91	57.39	18.07	81.93	34.79

(b) Results for ⟨verb, syntactic argument, theta role⟩ associations based on IGNGF cluster vs. FCA classification when using frame matching and probability estimates.

SRL method		**correct**			% of **incorrect**		%no label
		%total (R)	%labeled (P)	F	possible	impossible	
baseline		65.21	65.21	65.21	79.11	20.89	0.00
FCA (complete)	FM1, prob	57.23	59.80	58.48	80.46	19.54	4.30
IGNGF (basic)	FM, prob	62.39	64.07	63.22	29.66	70.34	2.64

Table 5.23: Overview of semantic role labeling results based on IGNGF cluster vs. FCA classification, based on the lexical resources only (ie. frame matching), (a) and on frame matching combined with probability estimates (b). When using frame matching only, both FCA and IGNGF achieved best results with a less detailed (partial) syntactic argument representation. The combined model performed best with the most detailed (complete) syntactic argument representation for FCA and with the simplest syntactic argument representation (basic) for IGNGF. We used each classification in the configuration it was shown to perform best: The labeling based on the FCA classification was computed using the less restrictive frame matching method (FM1) and the probability estimates are obtained from the SRL gold corpus (rather than the larger P7 corpus). Baseline is given by the default associations shown in Section 5.2.1.2.

FCA and IGNGF precision exceeded the baseline precision by roughly 5% and 7% respectively. For the labeling based on FCA a significantly larger proportion of the incorrect labels could have been corrected because the correct labels were at all associated with the corresponding instances: 82.05% vs. 18.07% (column *possible* in the table).

Finally, Table 5.23b shows the results when combining the labeling based on the clustering/classification with corpus probability estimates. The labelings were produced for each method with the configuration where it performed best. The table shows that results for the labelings based on the FCA classification were best when using the most detailed syntactic argument representation (complete), the less restrictive frame matching method (FM1) and the SRL gold corpus (*prob*, rather than the larger P7 corpus). The best labeling for IGNGF is obtained with the basic syntactic argument representation and also using the SRL gold corpus. For both labelings, based on the IGNGF clustering and the FCA classification, using probability estimates resulted in a gain in F-measure. The gain was of about 15 points for the FCA classification and almost 6 points for the IGNGF clustering. However the F-measure stayed below the baseline and precision, which for the labelings based on

frame matching only was above the baseline, decreased below the baseline precision. For both classifications a substantial percentage of instances could be labeled. The behaviour with regard to the incorrectly labeled instances was the same as when using frame matching only: For the FCA based labeling a large proportion (80.46%) of incorrectly labeled instances were also associated with the correct thematic roles, whereas for the IGNGF based labeling the proportion was of 29.66%. For the FCA classification the frame matching method and the corpus also had an impact (not shown in the table) on the results: The less restrictive frame matching method was most performant for all types of syntactic argument representations and for the simplest syntactic argument representation using all P7 instances to compute probability estimates produced the best F-measure.

To sum up, the IGNGF based labeling outperforms the one based on the FCA classification, both in terms of F-measure and number of labeled instances. Using the probability estimates increased the F-measure for both classifications. The size of this increase may be seen as an indication of how well the labeling can be adjusted to the corpus at hand. As this increase was more important for the FCA based method, this method is arguably more adjustable to the corpus to be labeled. Moreover, for both methods using probability estimates produced an important gain in the proportion of instances which could be labeled.

5.3 Conclusion and Discussion

In this chapter we assessed the ability of the classifications to associate groups of verbs with frames and thematic grids in two ways: On a global level we compared the associations induced by each classification with those found in a reference corpus. For a more fine grained assessment we performed a task based evaluation: We investigated whether and to what extent these classifications could support the labeling of ⟨verb, syntactic argument⟩ pairs in the reference corpus with thematic roles. While the first global level evaluation is more straight forward, the second type of evaluation on the syntax/semantics interface level involves various more complex steps which influence the labeling (ie. the evaluation result) and obfuscate the role of the classification.

We classified the roughly 4200 French verbs in the syntactic lexicon described in Section 3.1, using the FCA and IGNGF techniques. To build each classification we used the features which performed best for that classification. These were for the FCA classification the **scf-sem** feature set, consisting of subcategorisation frames and the additional semantic features shown in Table 3.5b. For the IGNGF method,

the feature set used was **grid-scf-sem**. It consisted of the **scf-sem** features and in addition the features derived from the translated verb classes.

The evaluation results obtained for the two classifications on the global and syntax/semantics interface level were complementary: In the global level evaluation the FCA classification performed better, whereas on the syntax/semantics interface the results were better for the IGNGF classification.

Global level evaluation. Regarding the global level evaluation, we found that both the IGNGF and FCA classifications covered an important number of ⟨verb, thematic grid⟩ corpus associations: This number was larger than for the classification proposed in [Sun *et al.*, 2010]. Of the ⟨verb, frame⟩ and ⟨verb, grid⟩ instances present in the corpus, a reasonable proportion was also present in the classifications: For frames, 51.58% for IGNGF and 56% for FCA and for grids, 48.11% for IGNGF and 76.90% for FCA. Thereby in this respect the FCA classification performed better than IGNGF: The associations of verbs with frames and grids engendered by the FCA classification seem to reflect the reference data better than the IGNGF classification. A possible reason is the overlapping character of the FCA classification. The difference in performance in this respect is also supported by the proportion of ⟨verb, frame⟩ and ⟨verb, grid⟩ pairs "lost" by the classifications: For IGNGF the proportion of reference pairs present in the original lexicon but not yielded by the classification was larger than for FCA, both for frames and grids (35.13% of frames and 48.11% of grids for IGNGF compared to 9.81% of frames and 11.95% of grids for FCA). Nevertheless, the ⟨verb, frame⟩ and ⟨verb, grid⟩ associations present in the reference and also produced by both classifications were roughly the same: 90% of the ⟨verb, frame⟩ pairs and 97% of the ⟨verb, grid⟩ present in the reference and in the IGNGF associations were also generated by the FCA classification. This suggests that FCA and IGNGF "agreed" to a large extent on what reference frames and grids to associate to the reference verbs.

With respect to the ⟨verb, grid⟩ associations which were obtained, for both classifications, by aligning verb groups with translated Verbnet classes, we found that often this alignment was not appropriate: Large verb classes were associated with small very specific translated Verbnet classes. This suggests that better ⟨verb, grid⟩ associations could be obtained using a better adapted alignment method. Another way to improve the verb and theta grid associations is by improving the translated Verbnet classes. We noticed that many English Verbnet classes have the same set of thematic roles but represent different semantic components (for instance the *hit-18.1* and *murder-42.1* Verbnet classes have the same thematic roles as the *cooking-45.3*

class, namely Agent, Instrument and Patient). Thus the classes grouped by thematic roles which we translate possibly blur semantic differences. We would obtain more accurate ⟨verb, grid⟩ associations by using translations of the original Verbnet classes and by retrieving the corresponding thematic role set after the alignment with the translated classes.

Evaluation on the syntax/semantics interface level. To evaluate the classifications on the syntax/semantics interface level we used them to label the reference corpus ⟨verb, syntactic argument⟩ instances with semantic roles and thus in effect built a simplified SRL system, following the method described in [Swier and Stevenson, 2005]. The resulting labeling is then compared with the reference annotations.

Both classifications performed reasonably well, with the IGNGF clustering systematically outperforming the FCA classifications: Based on the classifications only, the IGNGF SRL achieved an F-measure of 57.39 vs. 42.92 for FCA (compared to 76 in [Swier and Stevenson, 2005], with baseline 74). These results were below a default baseline of 65.21 where syntactic arguments are associated with default thematic roles. In addition, a large amount of instances could not be labeled. The reason for this in most cases was that the ⟨verb, syntactic argument⟩ combination was not engendered by the classification, rather than ambiguities in the associations with thematic roles. For IGNGF this is also supported by the relatively low proportion of reference ⟨verb, frame⟩ pairs covered (about 60% for IGNGF compared to almost 90% for FCA). In particular, the performance of both classifications for syntactic arguments involving the AOBJ and ATB functions was poor. A reason for this is that these are components of many frames in ⟨verb, frame⟩ pairs "lost" by both classifications (ie. present in the SRL gold and the lexicon, but not in the classification). In consequence, a promising way to improve the results would be to improve the association of verb groups with syntactic frames. For IGNGF a way to achieve this is suggested in Section 5.2.3: The first labeling results we obtained were low, but it was possible to improve them by adjusting the way subcategorisation frames and thematic role sets were associated to the clusters, ie. by adjusting the cluster labeling procedure. For FCA the association with subcategorisation frames may be improved by using attribute filtering indices (instead of just object filtering indices), as discussed in Section 5.2.2.

For the instances which could be labeled, both systems achieved a precision above the baseline, of 68% for FCA and 72.53% for IGNGF.

For both types of classifications the labeling with thematic roles was improved by combining the frame matching method with probability estimates, reaching an F-

measure of 62.82 for IGNGF and 58.48 for FCA (83 in [Swier and Stevenson, 2005]). All but about 4% of the reference instances could be labeled. Despite the improvements, these results are still below the default baseline (of 65.21). The improvements in F-measure when using probability estimates show that the associations produced with frame matching and the clustering/classification covered an important proportion of ⟨verb, syntactic argument, thematic role⟩ instances occuring in the reference corpus and these instances could successfully be used to predict further (correct) ⟨syntactic argument, thematic role⟩ combinations.

However, to further improve the results (and thus to complete and improve the resulting simple SRL system) we would need to account for such phenomena as polysemy and diathesis alternations for example by associating ⟨verb, syntactic argument⟩ instances with different thematic roles. Currently, the frame matching method employed does not support this: Using frame matching, every ⟨verb, syntactic argument⟩ pair can be labeled with at most one thematic role. Moreover, for a polysemous verb with more thematic grids, chances to find unambiguous ⟨verb, syntactic argument, thematic role⟩ associations are necessarily lower and thus in these cases ⟨verb, syntactic argument⟩ instances will more often remain unlabeled. Therefore this method ultimately is not suitable to evaluate the classifications with respect to handling polysemy. To improve it we would need to find a way to associate a ⟨verb, syntactic argument⟩ type with different thematic roles depending on the semantic context of the specific instance (token). In Verbnet this type of information is to a certain extent represented as selectional restrictions on the arguments or as semantic predicates, whereas our classifications don't provide this kind of information at all. However, the semantic features which were used and proved helpful to build both classifications do present analogies to some selectional restrictions in Verbnet.

Finally, the analysis of the associations obtained by combining frame matching with probability estimates further suggested that the amount of data (which in effect is used for learning ⟨syntactic argument, thematic role⟩ associations) may not be sufficient for the correct prediction of semantic roles.

These evaluation results show that the FCA and IGNGF classifications are complementary: On the global level the results suggest that the FCA classification, due to its overlapping character, better accounts for the verb polysemy phenomena present in the reference annotations. On the syntax/semantics interface level the results proved the IGNGF classification to perform better when used as a resource in the task of associating corpus ⟨verb, syntactic argument⟩ pairs with thematic roles. Ideally, one would like to combine the complementary strengths of both classifications,

for example to improve the semantic role labeling. This is not straight forward at this stage, however, since the frame matching method which we employed to produce the associations with thematic roles can not take into account verb polysemy or alternations (since it discards all ambiguous associations). A way to address this issue would be to enhance the frame matching procedure to consider semantic features extracted from the context of the ⟨verb, syntactic argument⟩ instances.

A further complementary feature of the two classifications is the following: If a verb in an IGNGF cluster is associated with a subcategorisation frame the verb is associated to the SCF by the lexicon only with some probability, whereas if the verb is associated to a subcategorisation frame by an FCA concept it is certain that it is asscociated to this SCF by the lexicon.

A shortcoming which the evaluation revealed to be common to both classifications is the association of verb groups with syntactic frames. Both classifications failed to associate verb groups with the correct syntactic frames – FCA often only provided the association with very frequent frames, whereas IGNGF often failed to produce the associations with these frequent frames. Therefore a promising way to improve the classifications is to focus on the association with syntactic frames. This is however hampered by the lack of reference resources to compare against (SRL gold is restricted in size and coverage).

To go more in depth with the comparison of the two classifications it would be interesting to compare the verb groups and associated sets of subcategorisation frames simultaneously. This could be done for example by applying ontology alignment and matching methods.

Chapter 6

Conclusion

6.1 Contributions

The main contribution of this thesis is the proposition of a novel approach to the automatic acquisition of a syntactic-semantic classification of French verbs, based on existing French and English lexical resources. Using this approach, we grouped more than 4200 French verbs into classes and associated these classes with both subcategorisation frames valid for the member verbs and thematic role sets representing the participants in the events described by these member verbs. The approach could be applied to corpus based data thus making the approach fully unsupervised and directly applicable to any language for which a parser is available

To build this classification we investigated and adapted two classification techniques which had not yet been used in this context: a symbolic classification method called Formal Concept Analysis (FCA) and IGNGF (Incremental Neural Gas with Feature Maximisation), an incremental neural (probabilistic) clustering method.

To relate the groups of verbs with thematic role sets, we used a novel translation approach where the verbs in English Verbnet classes are translated to French. This resulted in a resource where semantic groups of French verbs are mapped to thematic role sets of English Verbnet classes.

We performed a thorough evaluation of the resulting classifications first based on the groupings of verbs and second, based on the induced associations of verbs with syntactic frames and thematic grids.

In the following we detail each of the tasks addressed in this thesis, emphasising own contributions and highlighting differences to related studies.

Verb classification. While most approaches to the automatic acquisition of syntactic semantic verb classifications rely on features extracted from distributional data, the features our method is based on are extracted from available lexical resources. The two approaches are complementary. Whereas corpus based approaches permit building and tuning verb classes that are tailored to the corpus domain, approaches based on general-purpose lexical resources lay the basis for the construction of a general verb classification such as the English Verbnet, which represents the type of verb classification we aim at.

Most work on automatic verb classification concentrates on acquiring the verb classes that is, sets of verbs which are semantically and/or syntactically coherent. The specific features characterising that coherence are usually left implicit: they determine the clustering of similar verbs into verb classes but they do not explicitly label these classes. Our method explicits these associations with subcategorisation frames and thematic grids, and we propose an evaluation with respect to both the syntactic and semantic features associated to the verb groups.

Lastly, while there has been much work on automatically acquiring verb classes for English and to a lesser extent for German, Japanese and Italian, few studies have been conducted on the automatic classification of French verbs. In particular, to date the classifications we built are the first large scale Verbnet style classifications for French.

Classification methods. To our knowledge the FCA and IGNGF clustering methods have not to date been used to automatically acquire verb classifications. Using them for this application lead to a better understanding of the characteristics which are particularly supportive in this task.

One of the most distinctive features of FCA in the context of this application is its overlapping character, which, as we showed in Chapter 5.2, better accounts for verb polysemy. One of the most important drawbacks of this method are the large (unmanageable) number of concepts it produces. The resulting need to filter these concepts lead us to investigate the ability of the recently proposed concept selection indices to identify coherent concepts in the context of possibly noisy data. This investigation showed, that a linear combination of the stability and separation indices allowed to manage the complexity of the resulting concept lattice by selecting the concepts most relevant with respect to our data.

With respect to the IGNGF clustering method, its application clearly showed the benefits of the feature based metrics it employs both for producing the clusters and the associations with subcategorisation frames (cluster labeling).

Resources. The most important contribution of this thesis in terms of resources is the obtained large scale classification of French verbs, where groups of verbs are simultaneously associated with syntactic frames and semantic roles. Although the classification outcome is not a perfect syntactic-semantic verb classification, our experiments showed that the obtained classes were syntactically and semantically coherent and that they proved helpful in assigning thematic roles to verb arguments. Furthermore, our classification technique requires only minimal supervision and can easily be adapted to better suit a specific application.

Our method to build the verb classifications is based on a merged subcategorisation lexicon for French verbs and on translated English Verbnet classes, the development of which represents a further contribution of this thesis.

To evaluate the ⟨verb, frame, thematic grid⟩ associations engendered by the classifications we created a reference resource where we manually assigned thematic roles to instances of verbs and their syntactic arguments in a French treebank.

Evaluation. In most studies, classifications are evaluated based on the generated groups of verbs. In this thesis we also assessed the associations of verbs with syntactic frames and thematic grids by comparing them with a manually annotated reference corpus. Both evaluation schemes revealed commonalities and differences between the classifications obtained with the two classification techniques. First, the evaluation based on groups of verbs showed that for both classification methods, semantic features improved the classifications. We draw from the evaluation of ⟨verb, frame⟩ associations, that for both classification methods the association of the verb groups with frames was problematic. While FCA often associated groups of verbs only with very frequent and thus non-distinctive syntactic frames, these frames were often missed by the IGNGF method. Further work needs to focus on providing relevant ⟨verb, syntactic frame⟩ associations.

The evaluation of the thematic grid associations indicated the behaviour of the two classifications with respect to polysemy. They suggest that an overlapping or non-strict classification, as the one produced by FCA, better accounts for this phenomenon. However, the reference data sets used and our evaluation methodology currently do not permit a fine grained evaluation of this issue.

Finally, we assessed the syntactic and semantic coherence of our classifications in a task based evaluation. The verb classes were used as a resource in a procedure which makes use of Verbnet classes to assign thematic roles to verbs and their syntactic arguments, which is in essence a simplified SRL task. This evaluation allowed to estimate the capability of the generated verb classes at the syntax/semantics inter-

face, i.e. it allowed to assess whether the classes help generate the correct mapping between verb arguments and thematic roles. The classifications both proved to be helpful at this task, with the IGNGF classification clearly outperforming the FCA method. Although the results did not come close to state of the art scores in related work, this evaluation is a first step towards demonstrating the utility of the produced classifications in a task on "real-world" data.

6.2 Directions for Future Research

Semantic information. Our experiments showed that the kind of semantic information commonly referred to as "selectional restrictions" or "preferences" played an important role both in the acquisition of the verb classes and when using them to associate verb arguments with thematic roles. Therefore, an interesting question for future research is first how to extract and represent this information from available resources and second how to use it in the acquisition of the verb classes and in assigning thematic roles to verb arguments. In Section 3.1.1 we showed how characteristics of particular Verbnet classes and observations drawn from linguistic theories motivated the identification and extraction of some of the features used in the classification task. However, these features currently target only a limited number of classes. Their impact on the tasks explored here warrants a more thorough investigation.

[Mouton, 2010] proposes a French resource where words are grouped not only based on their belonging to the same lexical field but also on a notion of syntactic similarity, meaning that words in the same cluster occur in similar syntactic constructions. Since this can be considered a way of representing selectional preferences it would be interesting to explore its support in our verb class acquisition task[40].

A further type of semantic information is conveyed by the thematic role sets. In this thesis, they influence the nature and composition of the translated Verbnet classes, which in turn determine the association of verb groups with thematic role sets. There are many problems, both theoretical and empirical, with the traditional semantic role representation employed in this thesis [Samardžić, 2009]. One of these is for example which of the possibly infinite number of thematic roles to include into a finite role inventory, or how to group these roles to account for the fact that some thematic roles are more related than others. These problems also showed in our experiments: related roles were often confused and translated classes with the same

[40]The resource presented in [Sun and Korhonen, 2009] is similar in nature, but being for English is not relevant for this work.

thematic grid but distinct semantics (as for example *butter-9.9* and *put-9.1*) failed to clearly single out a class of verbs produced by the classifications. This points out the need to pay more attention to the impact of the role inventory: it is worthwhile to investigate which role inventory is more informative (to what task), how its roles are related in terms of their tendency to occur in the same syntactic environment and which roles tend to combine with which others. [Merlo and Van Der Plas, 2009] show some ways to pursue this direction.

The association of the verb classes with thematic roles also depends on the way the translated Verbnet classes are aligned with our classifications. Our experiments showed that our simple mapping procedure did not result in sufficiently accurate alignments. This suggests that applying more sophisticated alignment methods, possibly inspired by ontology mapping techniques could improve the results.

Classification methods. The classifications based on both methods, FCA and IGNGF, were less performant in associating the verb groups with relevant syntactic frames, it is therefore necessary to explore ways of improving this association. We presented some suggestions for achieving this in Section 5.1.1 and 5.1.2.

Another important question which needs to be addressed is how to adjust the classifications to properly take into account verb polysemy. FCA, due to its overlapping nature, does represent polysemy, but verbs seem to be attributed to too many classes, whereas IGNGF, being a crisp clustering method, always assigns verbs to exactly one class.

The FCA method, by its nature, produces hierarchical classifications. This is potentially beneficial for building verb classes in two ways. First, this structure is similar to the organisation of Verbnet verb classes. Second, previous research showed ([Li, 2008]) that the utility of verb classes in NLP applications strongly depends on the granularity of the classification. For some applications, as for example semantic role labeling systems, more coarse-grained classifications proved helpful, whereas in PP-attachment disambiguation only very-fine-grained verb class information was found to be useful. A hierarchical classification would possibly allow to choose an appropriate degree of granularity. Since FCA naturally builds hierarchical classifications, a profitable direction of future research is to explore the potential of the generated hierarchical organisation for tailoring the granularity of the verb classes. At first sight, IGNGF seems less suited for this task, but, being an iterative clustering method, possibly may be adjusted to produce a hierarchical structure.

A further interesting line of research is to investigate how the two clustering methods relate and to explore whether and how features useful to one method, can

prove efficient for the other.

Linguistic insights. Since our classifications relate groups of French verbs and their subcategorisation frames with the thematic roles of English Verbnet classes, they allow to identify differences and similarities in their syntactic realisation and behaviour in the two languages. An example of syntactic constructions which are more frequent in French than in English and which are captured by our classifications (Section 3.3.2), are pronominalisations, where the syntactic frames include a reflexive clitic pronoun (*Elle coiffe les cheveux de sa grand-mère/She does her grandmother's hair.* ↔ *Elle* ***se*** *coiffe/She does her hair.*[41]). Our qualitative discussion in Section 3.3.2 suggested that our classification correctly identified verb groups accepting these constructions. An interesting question is what English thematic roles and frames are realised in French by pronominalisations and whether and how the French reflexive pronouns in these constructions can/must be linked to thematic roles. However, in the French lexical resources it is often not clear whether a syntactic frame with a pronominalisation represents a genuine alternation rather than a syntactic frame of a verb which (in this meaning) must be used in a pronominal form. For example the construction (taken from Dicovalence) *L'incendie s'est communiqué à l'hôtel./The fire jumped over to the hotel.* is a syntactic frame of the lemma *se communiquer* rather than *communiquer*, because with this meaning this verb can be used only in the pronominal form. Unfortunately, this information can not be extracted from the Dicovalence lexical entry. Therefore, to efficiently use the classifications for exploring the use of these alternations in French, a further preprocessing and cleaning of the French lexical resources is needed.

[41] In English this particular alternation is expressed by a Patient role restricted to a body part (see for example *braid-41.2.2* Verbnet class)

Appendices

Appendix A

Frame Inventory of Merged Syntactic Lexicon

Table A.1: Frame inventory of the merged syntactic lexicon: unified frame representation, lexicon the frame was generated from and in parantheses the number of verbs it occured with in that lexicon. DV is short for Dicovalence, LA for Ladl tables and TL for TreeLex.

Frame representation	**Source information**
AOBJ:PP,ATB:XP,DUMMY:il	TL (1)
AOBJ:PP,DEOBJ:VPinf,DUMMY:il	TL (2)
AOBJ:PP,DUMMY:en,DUMMY:il	DV (1)
AOBJ:PP,DUMMY:il	TL (7)
AOBJ:PP,DUMMY:il,DUMMY:REFL	TL (1)
AOBJ:VPinf,DUMMY:il	TL (2)
AOBJ:VPinf,DUMMY:il,DUMMY:y	TL (1)
ATB:XP,DUMMY:il,DUMMY:REFL	TL (1)
DEOBJ:PP,DUMMY:il	DV (1), TL (4)
DEOBJ:PP,DUMMY:il,DUMMY:REFL	DV (3), TL (1)
DEOBJ:PP,DUMMY:il,DUMMY:y	TL (2)
DEOBJ:PP,POBJ:PP,DUMMY:en,DUMMY:il	DV (2)
DEOBJ:PP,POBJ:PP,DUMMY:il	TL (1)
DEOBJ:PP,POBJ:PP,DUMMY:il,DUMMY:REFL	DV (2)
DEOBJ:PP,POBJ:PP,POBJ:PP,DUMMY:en,DUMMY:il	DV (1)
DEOBJ:VPinf,DUMMY:il	TL (6)
DEOBJ:VPinf,DUMMY:il,DUMMY:REFL	DV (1), TL (1)
DEOBJ:VPinf,POBJ:PP,DUMMY:il,DUMMY:REFL	DV (1)
DUMMY:en,DUMMY:il	TL (1)
DUMMY:il	DV (1), LA (33), TL (11)
DUMMY:il,ATB:XP	LA (1)
DUMMY:il,DUMMY:REFL	TL (7)

Continued on next page

Table A.1 – continued from previous page

Frame representation	Source information
DUMMY:il,DUMMY:neg	LA (3)
DUMMY:il,DUMMY:y	TL (1)
OBJ:NP	TL (1)
OBJ:NP,AOBJ:PP,DUMMY:il	DV (4), TL (3)
OBJ:NP,AOBJ:PP,POBJ:PP,DUMMY:en,DUMMY:il	DV (1)
OBJ:NP,AOBJ:PP,POBJ:VPinf,DUMMY:en,DUMMY:il	DV (1)
OBJ:NP,ATB:XP	TL (1)
OBJ:NP,ATB:XP,DUMMY:il	TL (1)
OBJ:NP,DEOBJ:PP,DUMMY:il	DV (1), TL (8)
OBJ:NP,DEOBJ:PP,DUMMY:il,DUMMY:REFL	TL (1)
OBJ:NP,DEOBJ:VPinf,DUMMY:il	TL (1)
OBJ:NP,DUMMY:en,DUMMY:il,DUMMY:REFL	DV (1)
OBJ:NP,DUMMY:en,DUMMY:il,DUMMY:neg	DV (1)
OBJ:NP,DUMMY:il	DV (8), TL (12)
OBJ:NP,DUMMY:il,DUMMY:REFL	TL (5)
OBJ:NP,DUMMY:il,DUMMY:y	TL (4)
OBJ:NP,POBJ:PP,DUMMY:en,DUMMY:il,DUMMY:neg	DV (1)
OBJ:NP,POBJ:PP,DUMMY:il	TL (1)
OBJ:NP,POBJ:PP,DUMMY:il,DUMMY:y	TL (1)
OBJ:Ssub	DV (1), TL (1)
OBJ:Ssub,AOBJ:PP,DUMMY:il	TL (2)
OBJ:Ssub,ATB:XP,DUMMY:il	TL (3)
OBJ:Ssub,DEOBJ:PP,DUMMY:il	TL (2)
OBJ:Ssub,DUMMY:en,DUMMY:il,DUMMY:REFL	DV (1)
OBJ:Ssub,DUMMY:il	DV (2), TL (5)
OBJ:Ssub,DUMMY:il,DUMMY:REFL	DV (1), TL (3)
OBJ:Ssub,DUMMY:il,DUMMY:y	TL (1)
OBJ:Ssub,POBJ:PP,DUMMY:il	TL (1)
OBJ:VPinf,AOBJ:PP,ATB:XP,DUMMY:il	TL (1)
OBJ:VPinf,AOBJ:PP,DUMMY:il	DV (3), TL (3)
OBJ:VPinf,DEOBJ:PP,ATB:XP,DUMMY:il	TL (1)
OBJ:VPinf,DUMMY:il	DV (2), TL (6)
OBJ:VPinf,DUMMY:il,DUMMY:REFL	TL (3)
OBJ:VPinf,POBJ:PP,DUMMY:il	TL (1)
POBJ:PP,DUMMY:ca	DV (1)
POBJ:PP,DUMMY:en,DUMMY:il	TL (1)
POBJ:PP,DUMMY:il	TL (2)
POBJ:PP,DUMMY:il,DUMMY:REFL	TL (1)
POBJ:Ssub,DUMMY:il,DUMMY:REFL	TL (1)
SUJ:NP	DV (1875), LA (476), TL (644)
SUJ:NP,AOBJ:PP	DV (236), LA (71), TL (107)
SUJ:NP,AOBJ:PP,ATB:NP	DV (2)

Continued on next page

Table A.1 – continued from previous page

Frame representation	Source information
SUJ:NP,AOBJ:PP,ATB:XP	DV (2), TL (5)
SUJ:NP,AOBJ:PP,ATB:XP,DUMMY:REFL	TL (1)
SUJ:NP,AOBJ:PP,ATB:XP,DUMMY:les	LA (1)
SUJ:NP,AOBJ:PP,DEOBJ:PP	DV (5), TL (9)
SUJ:NP,AOBJ:PP,DEOBJ:PP,ATB:XP	TL (1)
SUJ:NP,AOBJ:PP,DEOBJ:PP,DUMMY:REFL	DV (3), TL (1)
SUJ:NP,AOBJ:PP,DEOBJ:PP,DUMMY:en,DUMMY:REFL	DV (1)
SUJ:NP,AOBJ:PP,DEOBJ:Ssub,DUMMY:REFL	DV (1)
SUJ:NP,AOBJ:PP,DEOBJ:VPinf	DV (1), TL (4)
SUJ:NP,AOBJ:PP,DEOBJ:VPinf,DUMMY:REFL	DV (2)
SUJ:NP,AOBJ:PP,DUMMY:REFL	DV (265), LA (18), TL (61)
SUJ:NP,AOBJ:PP,DUMMY:en	DV (7), LA (3), TL (5)
SUJ:NP,AOBJ:PP,DUMMY:en,DUMMY:REFL	TL (3)
SUJ:NP,AOBJ:PP,DUMMY:le,DUMMY:neg	LA (1)
SUJ:NP,AOBJ:PP,DUMMY:là	DV (1)
SUJ:NP,AOBJ:PP,DUMMY:neg	LA (2)
SUJ:NP,AOBJ:PP,DUMMY:neg,DUMMY:REFL	DV (1), LA (1)
SUJ:NP,AOBJ:PP,DUMMY:y	TL (1)
SUJ:NP,AOBJ:PP,POBJ:PP	DV (11), TL (4)
SUJ:NP,AOBJ:PP,POBJ:PP,DUMMY:REFL	DV (4), TL (1)
SUJ:NP,AOBJ:PP,POBJ:PP,DUMMY:en	DV (1)
SUJ:NP,AOBJ:PP,POBJ:Ssub	DV (1)
SUJ:NP,AOBJ:PP,POBJ:VPinf	DV (1)
SUJ:NP,AOBJ:PP,POBJ:VPinf,DUMMY:REFL	DV (1)
SUJ:NP,AOBJ:PP,POBJ:VPinf,DUMMY:en	DV (1)
SUJ:NP,AOBJ:Ssub	DV (20), LA (89), TL (1)
SUJ:NP,AOBJ:Ssub,DUMMY:REFL	DV (21), LA (46)
SUJ:NP,AOBJ:Ssub,DUMMY:en	LA (5)
SUJ:NP,AOBJ:Ssub,DUMMY:en,DUMMY:REFL	LA (4)
SUJ:NP,AOBJ:Ssub,DUMMY:neg,DUMMY:REFL	LA (1)
SUJ:NP,AOBJ:Ssub,POBJ:PP	LA (16)
SUJ:NP,AOBJ:Ssub,POBJ:PP,DUMMY:REFL	LA (2)
SUJ:NP,AOBJ:VPinf	DV (54), TL (37)
SUJ:NP,AOBJ:VPinf,DUMMY:REFL	DV (116), TL (22)
SUJ:NP,AOBJ:VPinf,DUMMY:en	DV (1), TL (2)
SUJ:NP,AOBJ:VPinf,DUMMY:neg,DUMMY:REFL	DV (1)
SUJ:NP,AOBJ:VPinf,POBJ:PP,DUMMY:REFL	DV (3)
SUJ:NP,AOBJ:VPinf,POBJ:VPinf,DUMMY:REFL	DV (1)
SUJ:NP,ATB:NP	DV (8)
SUJ:NP,ATB:NP,DUMMY:REFL	DV (21)
SUJ:NP,ATB:XP	DV (12), LA (30), TL (20)
SUJ:NP,ATB:XP,DUMMY:REFL	DV (43), LA (10), TL (23)

Continued on next page

Table A.1 – continued from previous page

Frame representation	Source information
SUJ:NP,ATB:XP,DUMMY:y	LA (1)
SUJ:NP,ATB:XP,DUMMY:y,DUMMY:REFL	TL (1)
SUJ:NP,DEOBJ:PP	DV (275), LA (99), TL (137)
SUJ:NP,DEOBJ:PP,ATB:NP	DV (1)
SUJ:NP,DEOBJ:PP,ATB:XP	DV (2), TL (5)
SUJ:NP,DEOBJ:PP,ATB:XP,DUMMY:REFL	DV (1), TL (2)
SUJ:NP,DEOBJ:PP,DUMMY:REFL	DV (414), TL (65)
SUJ:NP,DEOBJ:PP,DUMMY:en,DUMMY:REFL	DV (2)
SUJ:NP,DEOBJ:PP,DUMMY:en,DUMMY:neg	DV (1)
SUJ:NP,DEOBJ:PP,DUMMY:neg	DV (4)
SUJ:NP,DEOBJ:PP,DUMMY:neg,DUMMY:REFL	DV (1)
SUJ:NP,DEOBJ:PP,POBJ:PP	DV (32), TL (9)
SUJ:NP,DEOBJ:PP,POBJ:PP,ATB:XP	TL (1)
SUJ:NP,DEOBJ:PP,POBJ:PP,DUMMY:REFL	DV (19), TL (1)
SUJ:NP,DEOBJ:PP,POBJ:PP,DUMMY:neg	DV (1)
SUJ:NP,DEOBJ:PP,POBJ:PP,POBJ:PP	DV (1)
SUJ:NP,DEOBJ:PP,POBJ:VPinf	DV (1)
SUJ:NP,DEOBJ:Ssub	DV (22), LA (110), TL (6)
SUJ:NP,DEOBJ:Ssub,ATB:XP,DUMMY:REFL	LA (1)
SUJ:NP,DEOBJ:Ssub,DUMMY:REFL	DV (44), LA (60), TL (1)
SUJ:NP,DEOBJ:Ssub,DUMMY:en	LA (1)
SUJ:NP,DEOBJ:Ssub,DUMMY:en,DUMMY:neg	DV (1)
SUJ:NP,DEOBJ:Ssub,DUMMY:le,DUMMY:REFL	LA (1)
SUJ:NP,DEOBJ:Ssub,DUMMY:neg	LA (5)
SUJ:NP,DEOBJ:Ssub,DUMMY:neg,DUMMY:REFL	LA (2)
SUJ:NP,DEOBJ:Ssub,DUMMY:neg,DUMMY:y	LA (1)
SUJ:NP,DEOBJ:Ssub,POBJ:PP	DV (2), LA (56)
SUJ:NP,DEOBJ:Ssub,POBJ:PP,DUMMY:REFL	DV (4), LA (19)
SUJ:NP,DEOBJ:Ssub,POBJ:PP,DUMMY:en	LA (1)
SUJ:NP,DEOBJ:Ssub,POBJ:PP,DUMMY:en,DUMMY:REFL	LA (1)
SUJ:NP,DEOBJ:Ssub,POBJ:VPinf	DV (1)
SUJ:NP,DEOBJ:VPinf	DV (37), TL (28)
SUJ:NP,DEOBJ:VPinf,ATB:XP	TL (1)
SUJ:NP,DEOBJ:VPinf,ATB:XP,DUMMY:REFL	DV (1)
SUJ:NP,DEOBJ:VPinf,DUMMY:REFL	DV (127), TL (15)
SUJ:NP,DEOBJ:VPinf,DUMMY:en,DUMMY:neg	DV (1)
SUJ:NP,DEOBJ:VPinf,DUMMY:neg,DUMMY:REFL	DV (1)
SUJ:NP,DEOBJ:VPinf,DUMMY:y,DUMMY:REFL	TL (1)
SUJ:NP,DEOBJ:VPinf,POBJ:PP	DV (2), TL (1)
SUJ:NP,DEOBJ:VPinf,POBJ:PP,DUMMY:REFL	DV (3)
SUJ:NP,DUMMY:REFL	DV (2444), LA (99), TL (99)
SUJ:NP,DUMMY:REFL,DUMMY:y	LA (2)

Continued on next page

Table A.1 – continued from previous page

Frame representation	Source information
SUJ:NP,DUMMY:en	DV (9), LA (9), TL (5)
SUJ:NP,DUMMY:en,DUMMY:REFL	DV (9), LA (3), TL (5)
SUJ:NP,DUMMY:en,DUMMY:neg	DV (1)
SUJ:NP,DUMMY:le	DV (2), LA (6)
SUJ:NP,DUMMY:le,DUMMY:REFL	DV (1), LA (1)
SUJ:NP,DUMMY:le,DUMMY:neg	LA (1)
SUJ:NP,DUMMY:là	DV (1)
SUJ:NP,DUMMY:neg	DV (12), LA (8)
SUJ:NP,DUMMY:neg,DUMMY:REFL	DV (9), LA (5)
SUJ:NP,DUMMY:neg,DUMMY:y	LA (4)
SUJ:NP,DUMMY:y	LA (1), TL (1)
SUJ:NP,DUMMY:y,DUMMY:REFL	TL (1)
SUJ:NP,OBJ:NP	DV (4257), LA (2674), TL (1293)
SUJ:NP,OBJ:NP,AOBJ:PP	DV (451), LA (257), TL (178)
SUJ:NP,OBJ:NP,AOBJ:PP,ATB:XP	DV (1), LA (2), TL (1)
SUJ:NP,OBJ:NP,AOBJ:PP,DEOBJ:PP	DV (4), TL (5)
SUJ:NP,OBJ:NP,AOBJ:PP,DEOBJ:VPinf	TL (1)
SUJ:NP,OBJ:NP,AOBJ:PP,DUMMY:REFL	DV (1), TL (2)
SUJ:NP,OBJ:NP,AOBJ:PP,DUMMY:neg	DV (6), LA (2)
SUJ:NP,OBJ:NP,AOBJ:PP,DUMMY:y	TL (1)
SUJ:NP,OBJ:NP,AOBJ:PP,POBJ:PP	DV (5), TL (3)
SUJ:NP,OBJ:NP,AOBJ:PP,POBJ:VPinf	DV (1)
SUJ:NP,OBJ:NP,AOBJ:Ssub	DV (17), LA (201)
SUJ:NP,OBJ:NP,AOBJ:Ssub,DUMMY:neg	LA (2)
SUJ:NP,OBJ:NP,AOBJ:VPinf	DV (65), TL (32)
SUJ:NP,OBJ:NP,AOBJ:VPinf,DEOBJ:PP	TL (1)
SUJ:NP,OBJ:NP,AOBJ:VPinf,DUMMY:REFL	TL (2)
SUJ:NP,OBJ:NP,AOBJ:VPinf,DUMMY:neg	DV (1)
SUJ:NP,OBJ:NP,AOBJ:VPinf,POBJ:PP	DV (1)
SUJ:NP,OBJ:NP,AOBJ:VPinf,POBJ:VPinf	DV (1)
SUJ:NP,OBJ:NP,ATB:NP	DV (18)
SUJ:NP,OBJ:NP,ATB:PP	LA (61)
SUJ:NP,OBJ:NP,ATB:XP	DV (36), LA (14), TL (49)
SUJ:NP,OBJ:NP,ATB:XP,DUMMY:REFL	DV (1)
SUJ:NP,OBJ:NP,DEOBJ:PP	DV (389), LA (1286), TL (151)
SUJ:NP,OBJ:NP,DEOBJ:PP,ATB:XP	TL (1)
SUJ:NP,OBJ:NP,DEOBJ:PP,DUMMY:REFL	DV (3), TL (2)
SUJ:NP,OBJ:NP,DEOBJ:PP,POBJ:PP	DV (10)
SUJ:NP,OBJ:NP,DEOBJ:PP,POBJ:VPinf	DV (1)
SUJ:NP,OBJ:NP,DEOBJ:Ssub	DV (15), LA (217), TL (8)
SUJ:NP,OBJ:NP,DEOBJ:Ssub,DUMMY:neg	LA (2)
SUJ:NP,OBJ:NP,DEOBJ:Ssub,POBJ:PP	DV (1)

Continued on next page

Table A.1 – continued from previous page

Frame representation	Source information
SUJ:NP,OBJ:NP,DEOBJ:VPinf	DV (60), TL (20)
SUJ:NP,OBJ:NP,DEOBJ:VPinf,DUMMY:REFL	TL (1)
SUJ:NP,OBJ:NP,DUMMY:REFL	DV (284), LA (9), TL (48)
SUJ:NP,OBJ:NP,DUMMY:en	DV (1), LA (1)
SUJ:NP,OBJ:NP,DUMMY:le	LA (1)
SUJ:NP,OBJ:NP,DUMMY:là	LA (1)
SUJ:NP,OBJ:NP,DUMMY:neg	DV (8), LA (16)
SUJ:NP,OBJ:NP,DUMMY:neg,DUMMY:REFL	DV (1)
SUJ:NP,OBJ:NP,DUMMY:y	TL (5)
SUJ:NP,OBJ:NP,DUMMY:y,DUMMY:REFL	TL (1)
SUJ:NP,OBJ:NP,POBJ:PP	DV (745), LA (1387), TL (146)
SUJ:NP,OBJ:NP,POBJ:PP,ATB:NP	DV (1)
SUJ:NP,OBJ:NP,POBJ:PP,ATB:XP	DV (3), TL (3)
SUJ:NP,OBJ:NP,POBJ:PP,DUMMY:REFL	DV (6), TL (3)
SUJ:NP,OBJ:NP,POBJ:PP,DUMMY:neg	LA (1)
SUJ:NP,OBJ:NP,POBJ:PP,POBJ:PP	DV (5), LA (85)
SUJ:NP,OBJ:NP,POBJ:PP,POBJ:PP,POBJ:PP	DV (1)
SUJ:NP,OBJ:NP,POBJ:PP,POBJ:Ssub	DV (1)
SUJ:NP,OBJ:NP,POBJ:PP,POBJ:VPinf	DV (1)
SUJ:NP,OBJ:NP,POBJ:Ssub	DV (3), TL (1)
SUJ:NP,OBJ:NP,POBJ:Ssub,POBJ:PP	DV (1)
SUJ:NP,OBJ:NP,POBJ:Ssub,POBJ:Ssub	DV (1)
SUJ:NP,OBJ:NP,POBJ:Ssub,POBJ:VPinf	DV (1)
SUJ:NP,OBJ:NP,POBJ:VPinf	DV (17), LA (75), TL (4)
SUJ:NP,OBJ:NP,POBJ:VPinf,ATB:XP	DV (2)
SUJ:NP,OBJ:NP,POBJ:VPinf,POBJ:PP	DV (1)
SUJ:NP,OBJ:NP,POBJ:VPinf,POBJ:Ssub	DV (1)
SUJ:NP,OBJ:NP,POBJ:VPinf,POBJ:VPinf	DV (1)
SUJ:NP,OBJ:Ssub	DV (269), LA (459), TL (110)
SUJ:NP,OBJ:Ssub,AOBJ:PP	DV (85), LA (387), TL (12)
SUJ:NP,OBJ:Ssub,AOBJ:PP,ATB:XP	LA (1)
SUJ:NP,OBJ:Ssub,AOBJ:PP,DUMMY:neg	LA (3)
SUJ:NP,OBJ:Ssub,ATB:XP	DV (1), LA (22), TL (2)
SUJ:NP,OBJ:Ssub,DEOBJ:PP	DV (15), TL (9)
SUJ:NP,OBJ:Ssub,DEOBJ:PP,DUMMY:REFL	DV (1)
SUJ:NP,OBJ:Ssub,DEOBJ:Ssub	DV (1)
SUJ:NP,OBJ:Ssub,DUMMY:REFL	DV (79), LA (22), TL (12)
SUJ:NP,OBJ:Ssub,DUMMY:neg	DV (1), LA (9)
SUJ:NP,OBJ:Ssub,DUMMY:neg,DUMMY:REFL	LA (1)
SUJ:NP,OBJ:Ssub,DUMMY:y	TL (1)
SUJ:NP,OBJ:Ssub,POBJ:PP	DV (20), LA (199)
SUJ:NP,OBJ:Ssub,POBJ:PP,ATB:XP	LA (1)

Continued on next page

Table A.1 – continued from previous page

Frame representation	Source information
SUJ:NP,OBJ:VPinf	DV (161), TL (37)
SUJ:NP,OBJ:VPinf,AOBJ:PP	DV (69), TL (15)
SUJ:NP,OBJ:VPinf,AOBJ:PP,ATB:XP	TL (1)
SUJ:NP,OBJ:VPinf,AOBJ:PP,DEOBJ:PP	TL (1)
SUJ:NP,OBJ:VPinf,ATB:XP	DV (1), TL (3)
SUJ:NP,OBJ:VPinf,DEOBJ:PP	DV (3), TL (6)
SUJ:NP,OBJ:VPinf,DEOBJ:PP,DUMMY:REFL	TL (1)
SUJ:NP,OBJ:VPinf,DUMMY:REFL	DV (68), TL (4)
SUJ:NP,OBJ:VPinf,POBJ:PP	DV (9), TL (1)
SUJ:NP,OBJ:VPinf,POBJ:PP,DUMMY:REFL	TL (1)
SUJ:NP,POBJ:PP	DV (722), LA (665), TL (142)
SUJ:NP,POBJ:PP,ATB:XP	LA (2), TL (5)
SUJ:NP,POBJ:PP,ATB:XP,DUMMY:REFL	DV (5), LA (1)
SUJ:NP,POBJ:PP,ATB:XP,DUMMY:REFL,DUMMY:y	LA (1)
SUJ:NP,POBJ:PP,ATB:XP,DUMMY:y	LA (1)
SUJ:NP,POBJ:PP,DUMMY:REFL	DV (760), LA (222), TL (62)
SUJ:NP,POBJ:PP,DUMMY:REFL,DUMMY:y	LA (2)
SUJ:NP,POBJ:PP,DUMMY:en	DV (16), LA (11), TL (2)
SUJ:NP,POBJ:PP,DUMMY:en,DUMMY:REFL	DV (8), LA (5)
SUJ:NP,POBJ:PP,DUMMY:le	DV (2), LA (1)
SUJ:NP,POBJ:PP,DUMMY:neg	DV (1), LA (11)
SUJ:NP,POBJ:PP,DUMMY:neg,DUMMY:REFL	DV (1), LA (1)
SUJ:NP,POBJ:PP,DUMMY:y	LA (1)
SUJ:NP,POBJ:PP,POBJ:PP	DV (34)
SUJ:NP,POBJ:PP,POBJ:PP,DUMMY:REFL	DV (18)
SUJ:NP,POBJ:PP,POBJ:PP,DUMMY:en	DV (2)
SUJ:NP,POBJ:PP,POBJ:PP,DUMMY:le	DV (1)
SUJ:NP,POBJ:PP,POBJ:PP,POBJ:Ssub	LA (19)
SUJ:NP,POBJ:PP,POBJ:Ssub	DV (1)
SUJ:NP,POBJ:PP,POBJ:VPinf	LA (111)
SUJ:NP,POBJ:PP,POBJ:VPinf,DUMMY:REFL	LA (76)
SUJ:NP,POBJ:PP,POBJ:VPinf,DUMMY:en,DUMMY:REFL	LA (3)
SUJ:NP,POBJ:Ssub	DV (4), TL (1)
SUJ:NP,POBJ:Ssub,DUMMY:REFL	DV (2)
SUJ:NP,POBJ:Ssub,POBJ:PP	DV (1)
SUJ:NP,POBJ:Ssub,POBJ:Ssub	DV (1), LA (66)
SUJ:NP,POBJ:Ssub,POBJ:Ssub,DUMMY:REFL	LA (8)
SUJ:NP,POBJ:Ssub,POBJ:Ssub,DUMMY:neg	LA (1)
SUJ:NP,POBJ:VPinf	DV (6), LA (111), TL (4)
SUJ:NP,POBJ:VPinf,ATB:XP,DUMMY:REFL	DV (2)
SUJ:NP,POBJ:VPinf,DUMMY:REFL	DV (24), LA (76), TL (2)
SUJ:NP,POBJ:VPinf,DUMMY:en	DV (3)

Continued on next page

Table A.1 – continued from previous page

Frame representation	Source information
SUJ:NP,POBJ:VPinf,DUMMY:en,DUMMY:REFL	DV (2), LA (3)
SUJ:Ssub	DV (54), TL (3)
SUJ:Ssub,AOBJ:PP	DV (20)
SUJ:Ssub,AOBJ:PP,ATB:XP	DV (1)
SUJ:Ssub,AOBJ:PP,DUMMY:REFL	DV (7)
SUJ:Ssub,AOBJ:PP,POBJ:PP	DV (1)
SUJ:Ssub,AOBJ:PP,POBJ:Ssub	DV (1)
SUJ:Ssub,AOBJ:PP,POBJ:VPinf	DV (1)
SUJ:Ssub,AOBJ:Ssub	DV (4)
SUJ:Ssub,AOBJ:Ssub,DUMMY:REFL	DV (1)
SUJ:Ssub,AOBJ:VPinf	DV (6)
SUJ:Ssub,AOBJ:VPinf,DUMMY:REFL	DV (7)
SUJ:Ssub,ATB:XP	TL (1)
SUJ:Ssub,DEOBJ:PP	DV (4)
SUJ:Ssub,DEOBJ:PP,DUMMY:REFL	DV (8)
SUJ:Ssub,DEOBJ:Ssub	DV (1)
SUJ:Ssub,DEOBJ:Ssub,DUMMY:REFL	DV (2)
SUJ:Ssub,DEOBJ:VPinf	DV (1)
SUJ:Ssub,DEOBJ:VPinf,DUMMY:REFL	DV (5)
SUJ:Ssub,DUMMY:REFL	DV (65)
SUJ:Ssub,OBJ:NP	DV (154), LA (610), TL (3)
SUJ:Ssub,OBJ:NP,AOBJ:PP	DV (11)
SUJ:Ssub,OBJ:NP,AOBJ:PP,DUMMY:neg	DV (1)
SUJ:Ssub,OBJ:NP,AOBJ:Ssub	DV (3)
SUJ:Ssub,OBJ:NP,AOBJ:VPinf	DV (9)
SUJ:Ssub,OBJ:NP,ATB:XP	LA (2)
SUJ:Ssub,OBJ:NP,DEOBJ:PP	DV (8)
SUJ:Ssub,OBJ:NP,DEOBJ:Ssub	DV (2)
SUJ:Ssub,OBJ:NP,DEOBJ:VPinf	DV (5)
SUJ:Ssub,OBJ:NP,DUMMY:neg	DV (1), LA (2)
SUJ:Ssub,OBJ:NP,POBJ:PP	DV (10)
SUJ:Ssub,OBJ:Ssub	DV (10)
SUJ:Ssub,OBJ:Ssub,AOBJ:PP	DV (1)
SUJ:Ssub,OBJ:VPinf	DV (3), TL (1)
SUJ:Ssub,OBJ:VPinf,AOBJ:PP	DV (1)
SUJ:Ssub,POBJ:PP	DV (3), LA (144)
SUJ:Ssub,POBJ:PP,ATB:XP	LA (4)
SUJ:Ssub,POBJ:PP,ATB:XP,DUMMY:REFL	LA (1)
SUJ:Ssub,POBJ:PP,DUMMY:REFL	DV (5), LA (11)
SUJ:Ssub,POBJ:PP,DUMMY:en	LA (2)
SUJ:Ssub,POBJ:PP,DUMMY:neg	LA (2)
SUJ:Ssub,POBJ:Ssub	DV (1)

Continued on next page

Table A.1 – continued from previous page

Frame representation	Source information
SUJ:Ssub,POBJ:VPinf	DV (1)
SUJ:VPinf	DV (61), TL (3)
SUJ:VPinf,AOBJ:PP	DV (27), TL (1)
SUJ:VPinf,AOBJ:PP,ATB:XP	DV (1)
SUJ:VPinf,AOBJ:PP,DUMMY:REFL	DV (4)
SUJ:VPinf,AOBJ:PP,DUMMY:en	DV (1)
SUJ:VPinf,AOBJ:PP,POBJ:PP	DV (1)
SUJ:VPinf,AOBJ:PP,POBJ:Ssub	DV (1)
SUJ:VPinf,AOBJ:PP,POBJ:VPinf	DV (1)
SUJ:VPinf,AOBJ:Ssub	DV (4)
SUJ:VPinf,AOBJ:Ssub,DUMMY:REFL	DV (2)
SUJ:VPinf,AOBJ:VPinf	DV (5), TL (1)
SUJ:VPinf,AOBJ:VPinf,DUMMY:REFL	DV (4)
SUJ:VPinf,ATB:XP	TL (2)
SUJ:VPinf,DEOBJ:PP	DV (3), TL (1)
SUJ:VPinf,DEOBJ:PP,DUMMY:REFL	DV (8)
SUJ:VPinf,DEOBJ:VPinf	DV (3), TL (1)
SUJ:VPinf,DEOBJ:VPinf,DUMMY:REFL	DV (5)
SUJ:VPinf,DUMMY:REFL	DV (50)
SUJ:VPinf,OBJ:NP	DV (135), TL (4)
SUJ:VPinf,OBJ:NP,AOBJ:PP	DV (13)
SUJ:VPinf,OBJ:NP,AOBJ:PP,DUMMY:neg	DV (1)
SUJ:VPinf,OBJ:NP,AOBJ:Ssub	DV (5)
SUJ:VPinf,OBJ:NP,AOBJ:VPinf	DV (7)
SUJ:VPinf,OBJ:NP,DEOBJ:PP	DV (11)
SUJ:VPinf,OBJ:NP,DEOBJ:PP,POBJ:PP	DV (1)
SUJ:VPinf,OBJ:NP,DEOBJ:PP,POBJ:VPinf	DV (1)
SUJ:VPinf,OBJ:NP,DEOBJ:VPinf	DV (6)
SUJ:VPinf,OBJ:NP,DUMMY:REFL	DV (2)
SUJ:VPinf,OBJ:NP,DUMMY:neg	DV (1)
SUJ:VPinf,OBJ:NP,POBJ:PP	DV (7)
SUJ:VPinf,OBJ:NP,POBJ:VPinf	DV (1)
SUJ:VPinf,OBJ:Ssub	DV (4), TL (1)
SUJ:VPinf,OBJ:VPinf	DV (2), TL (3)
SUJ:VPinf,OBJ:VPinf,AOBJ:PP	DV (2)
SUJ:VPinf,OBJ:VPinf,DUMMY:REFL	DV (1)
SUJ:VPinf,POBJ:PP	DV (2)
SUJ:VPinf,POBJ:PP,DUMMY:REFL	DV (2)
SUJ:VPinf,POBJ:Ssub	DV (1)
SUJ:VPinf,POBJ:VPinf	DV (1)

abattre	SUJ:NP,POBJ:PP
acheter	SUJ:NP,OBJ:NP,ATB:XP
aller	SUJ:NP,AOBJ:PP,DEOBJ:PP
annoncer	SUJ:NP,OBJ:VPinf
atténuer	SUJ:NP
circuler	SUJ:NP,AOBJ:PP,DEOBJ:PP
concevoir	SUJ:NP,OBJ:NP,ATB:XP
consolider	SUJ:NP,OBJ:NP,POBJ:Ssub
contribuer	SUJ:NP,OBJ:NP
contribuer	SUJ:NP,OBJ:NP,AOBJ:PP
corriger	SUJ:NP,OBJ:Ssub
courir	SUJ:NP,DEOBJ:PP,POBJ:PP
décrire	SUJ:NP,OBJ:NP,POBJ:PP
déplacer	SUJ:NP,OBJ:NP,AOBJ:PP,DEOBJ:PP
déposer	SUJ:NP,OBJ:NP,AOBJ:PP
empiler	SUJ:NP,POBJ:PP
examiner	SUJ:NP,OBJ:NP,AOBJ:PP
former	SUJ:NP,AOBJ:PP
former	SUJ:NP,OBJ:NP,ATB:XP
frapper	SUJ:NP,AOBJ:PP
gronder	SUJ:NP,POBJ:PP
gémir	SUJ:NP,OBJ:Ssub
heurter	SUJ:NP,AOBJ:PP,ATB:XP
installer	SUJ:NP,AOBJ:PP
installer	SUJ:NP,OBJ:NP,AOBJ:PP
montrer	SUJ:NP,ATB:XP
mêler	SUJ:NP,AOBJ:PP
mêler	SUJ:NP,POBJ:PP
naviguer	SUJ:NP,AOBJ:PP
osciller	SUJ:NP,AOBJ:PP,DEOBJ:PP
participer	SUJ:NP,AOBJ:PP,POBJ:PP
porter	SUJ:NP,OBJ:NP,AOBJ:PP,DEOBJ:PP
regarder	SUJ:NP,OBJ:VPinf
ressentir	SUJ:NP,OBJ:NP,ATB:XP
ressentir	SUJ:NP,OBJ:NP,POBJ:PP
retrancher	SUJ:NP,POBJ:PP
répartir	SUJ:NP,POBJ:PP
révéler	SUJ:NP,ATB:XP
révéler	SUJ:NP,DEOBJ:PP
révéler	SUJ:NP,POBJ:PP
tressaillir	SUJ:NP
voir	SUJ:NP,OBJ:VPinf

Table A.2: Verb, frame instances (42 types) present in annotated corpus but not in merged lexicon.

Appendix B

FCA Evaluation

B.1 Evaluating ⟨verb, frame⟩ associations.

Table B.1 shows the ⟨verb, frame⟩ pairs present in the SRL gold and the lexicon but not in the classification (see Section 5.2.2.1).

Table B.1: Verb, frame pairs (31 types) present in reference corpus annotations but not in the FCA classification.

Verb	**SCF**
acheter	SUJ:NP,AOBJ:PP
adresser	SUJ:NP,AOBJ:PP
adresser	SUJ:NP,OBJ:NP,AOBJ:PP
affirmer	SUJ:NP,OBJ:VPinf
aller	SUJ:NP,AOBJ:PP
aller	SUJ:NP,OBJ:VPinf
aller	SUJ:NP,POBJ:VPinf
associer	SUJ:NP,OBJ:NP,AOBJ:PP,POBJ:PP
collaborer	SUJ:NP,AOBJ:PP
concourir	SUJ:NP,AOBJ:PP
contribuer	SUJ:NP,AOBJ:PP
contribuer	SUJ:NP,AOBJ:VPinf
coopérer	SUJ:NP,AOBJ:PP
dire	SUJ:NP,OBJ:NP,ATB:XP
dire	SUJ:NP,OBJ:VPinf
définir	SUJ:NP,OBJ:NP,ATB:XP
envisager	SUJ:NP,OBJ:VPinf
garder	SUJ:NP,OBJ:NP,ATB:XP
heurter	SUJ:NP,AOBJ:PP
incorporer	SUJ:NP,OBJ:NP,AOBJ:PP

Continued on next page

Table B.1 – continued from previous page

Verb	SCF
maintenir	SUJ:NP,OBJ:NP,ATB:XP
participer	SUJ:NP,AOBJ:PP
passer	SUJ:NP,AOBJ:PP,DEOBJ:PP
percevoir	SUJ:NP,OBJ:NP,ATB:XP
porter	SUJ:NP,AOBJ:PP
regarder	SUJ:NP,AOBJ:PP
renvoyer	SUJ:NP,AOBJ:PP
répliquer	SUJ:NP,AOBJ:PP
répondre	SUJ:NP,AOBJ:PP
voir	SUJ:NP,ATB:XP
voir	SUJ:NP,OBJ:Ssub,ATB:XP

In Table B.2 we list the ⟨verb, theta grid⟩ pairs present in the SRL, which were not compatible with any theta grid associated to the verb by the classification (see Section 5.2.2.1).

Table B.2: Verb, theta grid instances (38 types) present in reference corpus annotations but not compatible with any thematic grid associated to the verb by the FCA classification.

Verb	theta grid
affirmer	Cause-Theme
aller	End-Start-Theme
aller	End-Theme
appréhender	Agent-Topic
certifier	Agent-Topic
concourir	End-Theme
conserver	Agent-Beneficiary-Theme
contribuer	Agent-Theme
contribuer	Agent-ThemeSym
contribuer	Agent-ThemeSym-ThemeSym
contribuer	ThemeSym-ThemeSym
corriger	Instrument-Theme
courir	End-Start-Theme
dire	Agent-Theme-Topic
désigner	Agent-PredAtt-Theme
désigner	Cause-Topic
désigner	PredAtt-Theme
empiler	Location-Theme
examiner	Experiencer-Theme
former	Agent-PredAtt-Theme

Continued on next page

Table B.2 – continued from previous page

Verb	theta grid
fortifier	Agent-Instrument-Patient
incorporer	Agent-PatientSym-PatientSym
inonder	Agent-Instrument-Theme
inonder	Instrument-Theme
lancer	Instrument-Theme
osciller	End-Start-Theme
participer	End-Theme
prendre	Agent-PredAtt-Theme
préserver	Agent-Start-Theme
préserver	Instrument-Theme
ressentir	Experiencer-Theme
répondre	Theme-Topic
réserver	Agent-Beneficiary-Theme
saisir	Agent-Theme-Topic
saisir	Experiencer-Topic
scruter	Experiencer-Theme
supprimer	Cause-Theme
écouter	Experiencer-Theme

B.2 Evaluation on the syntax/semantics interface level.

In Section 5.2.2.2 we assess to what degree the verb classification automatically acquired with FCA supports the automatic assignment of thematic roles to verb syntactic arguments. This evaluation is carried out by using the classification in a simplified semantic role labeling task as proposed in [Swier and Stevenson, 2005]. The most relevant aspects of this evaluation are discussed in Section 5.2.2.2, here we analyse the performance of our method (based on the FCA classification) in terms of:

Semantic roles What is the performance for each thematic role?

Syntactic arguments What is the performance for each syntactic argument?

Number of subcategorisation frames In this set of experiments we organise the verbs in groups depending on the number of the subcategorisation frames they are associated with by the FCA classification and evaluate the associated thematic roles by these groupings.

Polysemy classes We consider that a verb can be polysemous in two ways. The polysemy class may be

1) the number of FCA concepts it is a member of or
2) the number of translated Verbnet classes it is a member of.

Here we evaluate the results for each polysemy class.

Our previous results suggested that the more restrictive frame matching method produces a more accurate labeling, but it allows too few distinct syntactic arguments to be labeled and thus can hardly be adapted using corpus data. In contrast the less restrictive frame matching method produced a labeling which was somewhat less accurate but allowed for an important improvement when using probability estimates. In the following we will therefore use the less restrictive frame matching method.

Evaluation by semantic role. Table B.3a shows the results of the labeling obtained by using the frame and argument matcher only, per semantic role and by decreasing F-measure. A first observation is that only three thematic roles could unambiguously be mapped to the corpus instances: Agent, Patient and Theme, whereas the thematic roles used in the annotation were the following (in parantheses the number of corpus instances labeled with the corresponding thematic role): Agent (1263), Beneficiary (25), Cause (27), End (302), Experiencer (59), Extent (2), Instrument (79), Location (48), Patient (261), PredAtt (141), Start (60), Theme (1106), Topic (237). For the assigned thematic roles the number of instances for which the correct role is impossible for the system to detect is relatively low.

The results shown in Table B.3a only depend on the FCA classification which was built automatically from lexical resources and are therefore independent of corpus data. Table B.3b shows the results per role, again by decreasing F-measure when combining the labeling produced with the frame and argument matcher with corpus data. ⟨verb, syntactic argument⟩ instances in the corpus are labeled using the unique associations obtained using the FCA classifications and then the frequency of obtained ⟨syntactic argument, semantic role⟩ associations is used to label further, non-unique ⟨verb, syntactic argument⟩ associations. We saw that there were only three thematic roles (Agent, Theme, Patient) which could be uniquely associated to annotated ⟨verb, syntactic argument⟩ instances. For all of these roles, using probability estimates increased the F-measure by ≈27, ≈40 and ≈9% respectively and the number of labeled instances could be increased significantly. As with most SRL systems, the labeling of the Agent role was most accurate with and without using

(a) SRL using frame and argument matcher only.

Role	F	R	P	total	labeled	correct	impossible	not labeled
Agent	72.16	56.53	99.72	1263	716	714	0	547
Theme	29.70	19.00	67.96	1105	309	210	28	796
Patient	8.12	4.98	22.03	261	59	13	3	202
Overall	37.94	25.99	70.24	3605	1334	937	91	2271

(b) SRL using the frame and argument matcher combined with frequency information from the P7 corpus.

Role	F	R	P	total	labeled	correct	impossible	not labeled
Agent	99.12	98.65	99.60	1263	1251	1246	3	12
Theme	71.42	69.77	73.15	1105	1054	771	82	51
Patient	17.37	17.24	17.51	261	257	45	3	4
Overall	58.45	57.20	59.77	3605	3450	2062	272	155

Table B.3: Performance of the semantic role labeling per role and by descendig F-measure. The SRL is obtained using the frame and argument matcher only (a) or combined with frequency information from the corpus (b).

Legend:

F:	F-measure
R:	recall
P:	precision
total:	total number of instances labeled with this role in reference.
labeled:	total number of instances labeled with this role by the system.
correct:	number of instances correctly labeled with this role by the system.
impossible:	number of instances labeled in the reference with this role, where correct role not in FCA assoc
not labeled:	number of instances labeled in the reference with this role but not labeled by the system.

probability estimates. Labeling the Theme role was less accurate when using the system based on the FCA classification only (29.70) showing that for this role the FCA classification and the frame matching method provide little unambiguous mappings. However, using the corpus data it was possible to achieve a good F-measure of 71.42 points which shows that many of the associations for this role provided by the FCA classification and frame matching corresponded to the realisations of this role in the corpus.

Figure B.1 shows the difference in the labeling results (F-measure) when using the frame and argument matching only compared to the SRL system combining frame and argument matching with corpus frequency data. The roles are listed by decreasing difference in f-measure for the combined system vs. the system using frame and argument matching only.

The F-measure for the Patient role was low for the system based on FCA only but could be doubled using probability estimates. This suggests that, despite being rather inaccurate, the associations for this role provided by the FCA classification still were correct for many instances occuring in the corpus.

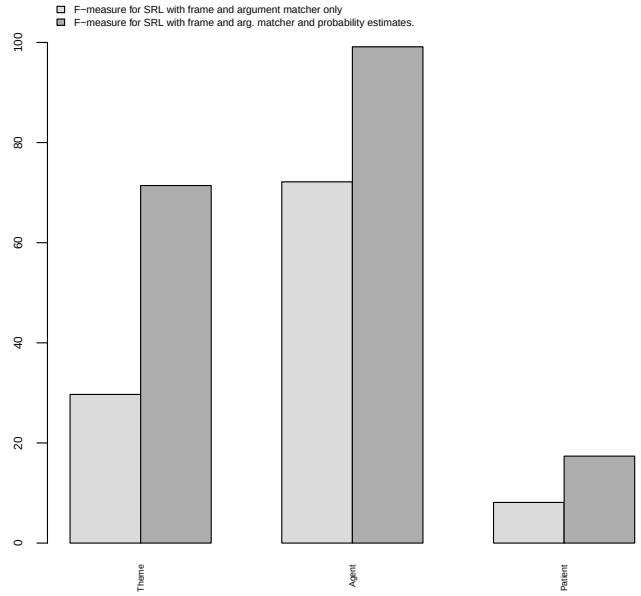


Figure B.1: F-measure for SRL obtained by frame and argument matching only vs. frame and argument matching combined with corpus frequency data, by thematic roles.

Evaluation by syntactic arguments. In this section we further investigate the performance of the two systems, one based on lexical resources only, the other on combining information from lexical resources with corpus data. Here we check the F-measure for each of the syntactic arguments occuring in the corpus and associated with a thematic role by our systems.

Instances with the following syntactic arguments could be labeled in the frame matching step: AOBJ:PP, DEOBJ:PP, OBJ:NP, OBJ:Ssub, POBJ:PP, SUJ:NP compared to a number of 12 in the reference.

Table B.4 shows the results per syntactic argument, by decreasing F-measure, first for the semantic role labeling obtained using the frame and argument matcher only (Table B.4a), second using the frame and argument matcher combined with corpus frequency information (Table B.4b).

The system based on the FCA classification only produced the best labeling for the SUJ:NP and OBJ:NP syntactic arguments.

Figure B.2 shows the difference in the labeling results (F-measure) when using the frame and argument matching only compared to the SRL system combining frame and argument matching with corpus frequency data. The syntactic arguments

(a) SRL using frame and argument matcher

Role	F	R	P	total	labeled	correct	impossible	not labeled
SUJ:NP	58.45	44.26	86.04	1602	824	709	41	778
OBJ:NP	29.67	19.15	65.87	1149	334	220	8	815
DEOBJ:PP	5.06	3.12	13.33	64	15	2	2	49
POBJ:PP	3.40	2.25	6.94	222	72	5	24	150
OBJ:Ssub	1.14	0.65	4.76	155	21	1	0	134
Overall	37.94	25.99	70.24	3605	1334	937	91	2271

(b) SRL using frame and argument matcher combined with corpus frequency information.

Role	F	R	P	total	labeled	correct	impossible	not labeled
SUJ:NP	77.86	77.47	78.25	1602	1586	1241	46	16
OBJ:NP	67.19	66.75	67.64	1149	1134	767	45	15
DEOBJ:PP	4.84	4.69	5.00	64	60	3	15	4
POBJ:PP	8.11	8.11	8.11	222	222	18	33	0
OBJ:Ssub	0.65	0.65	0.65	155	154	1	0	1
Overall	58.45	57.20	59.77	3605	3450	2062	272	155

Table B.4: Performance of the semantic role labeling per syntactic argument and by decreasing F-measure. The SRL is obtained using the frame and argument matcher only (a) or combined with P7 frequency information (b).

Legend:

R:	recall
F:	F-measure
P:	precision
total:	total number of instances labeled with this role in reference.
labeled:	total number of instances labeled with this role by the system.
correct:	number of instances correctly labeled with this role by the system.
impossible:	number of instances labeled in the reference with this role, where correct role not in FCA assoc
not labeled:	number of instances labeled in the reference with this role but not labeled by the system.

are listed by decreasing difference in f-measure for the combined system vs. the system using frame and argument matching only. According to this figure we can see that there is an important increase in F-measure when also using probability estimates and that this increase is more important in the case of OBJ:NP. This is not surprising because OBJ:NP is a syntactic argument which is frequent and may be the realisation of a larger number of thematic roles and thus the need of corpus data to disambiguate them is foreseeable. The results do not seem helpfull in the case of the other labeled syntactic arguments: There is a small increase in F-measure for POBJ:PP but none for DEOBJ:PP and OBJ:Ssub. An explanation for this may be that DEOBJ:PP and OBJ:Ssub are the less frequent among the labeled syntactic arguments.

In the following we built verb classes based on the following criteria:

1. Number of subcategorisation frames a verb has,

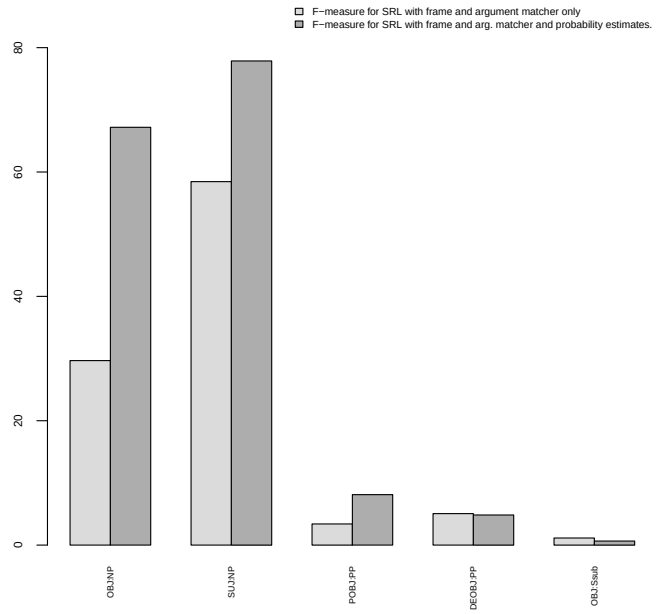


Figure B.2: F-measure for SRL obtained by frame and argument matching only vs. frame and argument matching combined with corpus frequency data, by thematic roles.

2. Polysemy, ie. the number of classes a verb was a member of.

   (a) Number of FCA classes the verb was a member of,

   (b) Number of translated classes the verb is a member of.

For each of these class sets we analyse:

- the F-measures per class for the SRL based on lexical resources only,
- the F-measures per class for the SRL combining lexical resources and corpus data,
- the improvement in F-measure for the combined system compared to the system based on lexical data only.

We could not find a clear correlation between the labeling results (based on lexical and corpus data) and the class membership for any of these class sets.

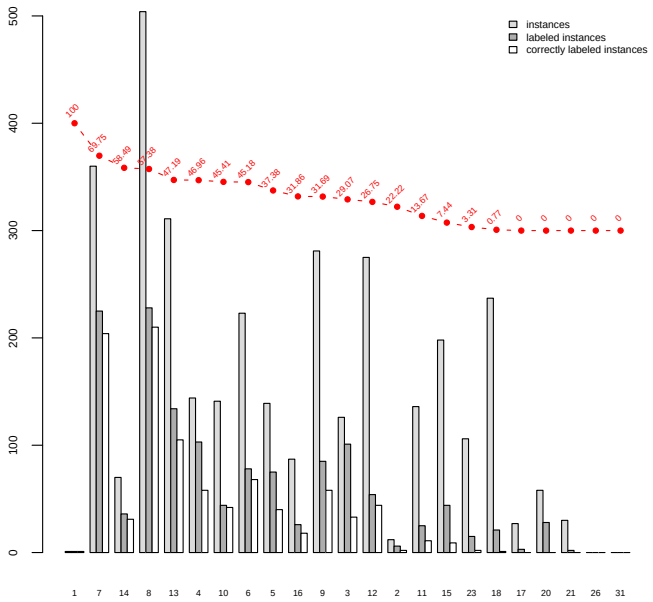
Evaluation by number of syntactic frames. In the following we analyse the number of correctly labeled instances dependent on the number of subcategorisation frames the verbs are associated with by the lexicon.

Figure B.3 shows for each class of verbs with a given number of subcategorisation frames (i) the number of annotated instances (light grey), (ii) the number of annotated instances labeled by the SRL system (darker grey), (iii) the number of correctly labeled instances (white) and (iv) the f-measure for this class (red lines and figures) for the SRL based on lexical data only (Figure B.3a) and on lexical and corpus data combined (red lines and figures, Figure B.3b). The figure shows that for most of these scf-classes the F-measure is improved when using the probability estimates. In addition, instances with verbs with 26 and 31 subcategorisation frames could not be labeled using frame matching but a labeling was possible with probability estimates from corpus data. In Figure B.4 we show the difference in F-measure of the SRL based on lexical data only (light grey bar) compared to the SRL using lexical and corpus data combined. The classes are ordered by decreasing difference in F-measure, so the difference is biggest for the leftmost class. This graphic shows that for most classes combining the lexical data with corpus data improved the F-measure of the labeling. The improvement was most important for verbs in 20 FCA classes (0 to 80%) and was in general larger for the verbs with many subcategorisation frames. This result is plausible, considering that for these verbs ⟨verb, syntactic argument⟩ instances the probability to be associated with more thematic roles is greater. In contrast, for verbs with 1 or 2 subcategorisation frames there is no improvement in F-measure, the F-measure even decreases slightly. The verbs with 14 subcategorisation frames don't seem to fit into this pattern.

Overall we think that this data does not suggest an obvious correlation between the number of subcategorisation frames a verb is associated with and the semantic role labeling results.

Evaluation by Polysemy In this section we analyse the performance of the semantic role labeling in dependence of the polysemy classes of the verbs. In the first analysis the polysemy class of a verb is the number of thematic role sets it is associated with in the FCA classification. Figure B.5 shows the distribution of verbs in the various polysemy classes. Figures B.5a and B.5b show the results of the semantic role labeling based on the FCA classification only and on the combined lexical and corpus data respectively. The x-axis gives the polysemy classes (by decreasing f-measure for the labeling based on the FCA classification) and the y-axis the number of annotated instances in the corresponding class. For each of the polysemy classes

(a) SRL using frame and argument matcher only



(b) SRL using frame and argument matcher combined with corpus frequency information.

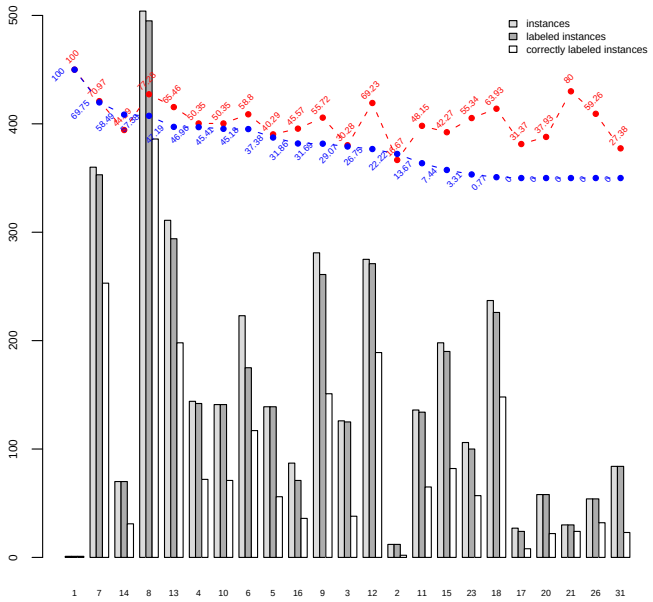


Figure B.3: SRL performance by number of subcategorisation frames.

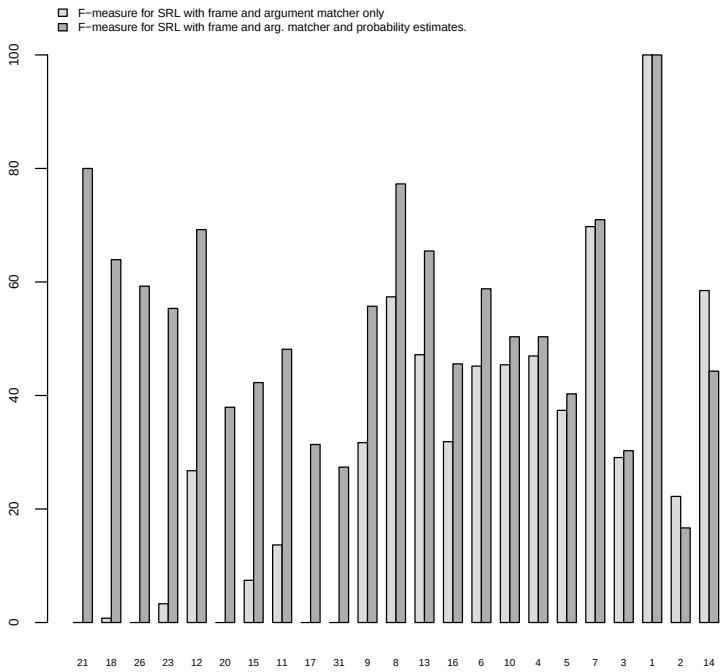
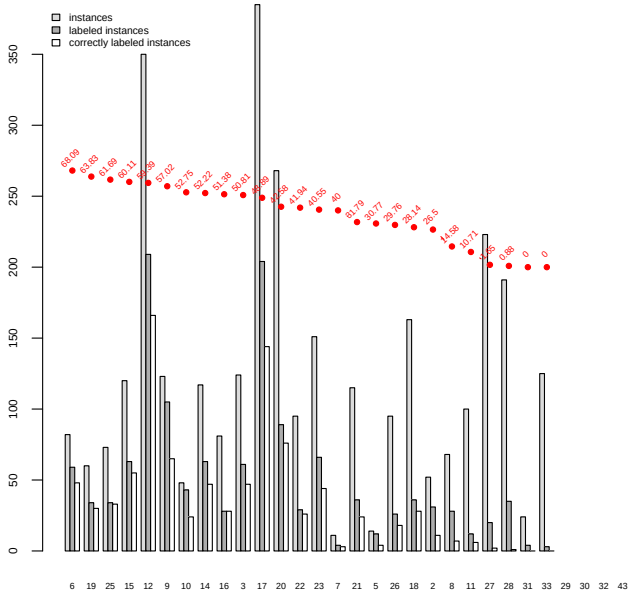


Figure B.4: Difference in F-measure for SRL using frame and argument matching only vs. SRL combining lexical and corpus data, by number of subcategorisation frames per verb.

the figure shows (i) the number of annotated instances (light grey bars), (ii) the number of instances which could be labeled by the SRL system (darker grey bar), (iii) the number of correctly labeled instances (white bar) and (iv) the F-measure (red line and figures). In Figure B.7 we added the F-measures for the labeling based on lexical data only for ease of comparison (blue line and blue numbers).

The figure shows that in our data the verbs were associated with at least 2 and at most 43 thematic role sets and there were 30 polysemy classes. For the labeling based on the FCA classification only, the F-measure per polysemy class ranged from 0 to 68.09 (red line and numbers) and the best F-measure was obtained for verbs which were in 6 classes (Figure B.5a). For the labeling based on the combination of the FCA classification and corpus data the F-measure ranged from 11.76 to 92.44. In general, for instances with verbs in a large number of FCA classes the performance of the system based on the FCA classification only was low (the rightmost bars in Figure B.5b) but could be improved using corpus data to F-measures ranging

(a) SRL using frame and argument matcher



(b) SRL using frame and argument matcher combined with corpus frequency information.

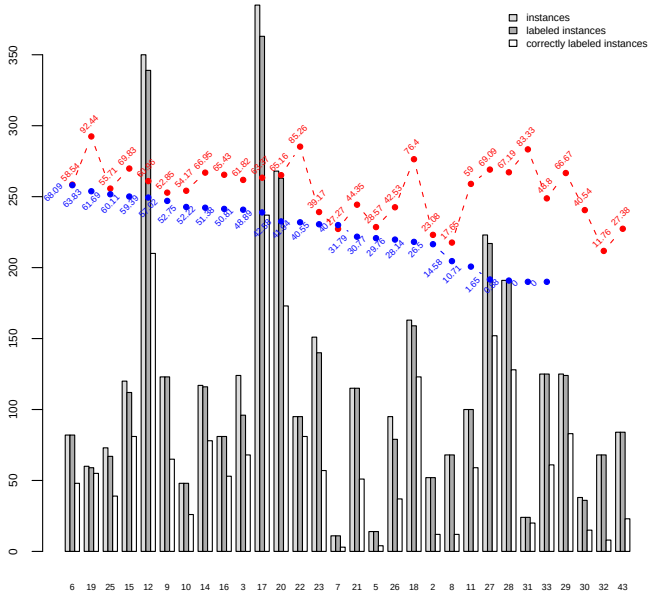


Figure B.5: SRL performance by number of FCA classes a verb is a member of.

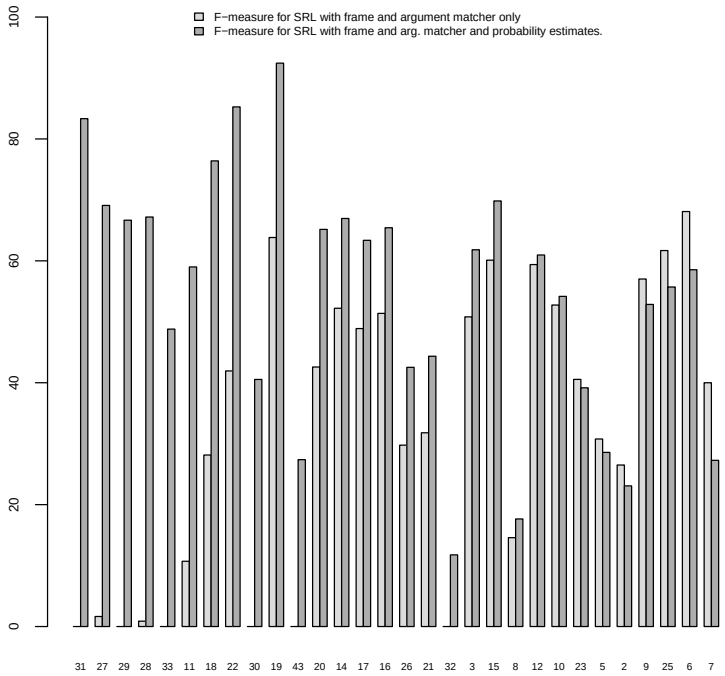


Figure B.6: Difference in F-measure between labeling based on the FCA classification only compared to labeling using a combination of FCA classification and corpus data.

from 11.76 to 83.33. In Figure B.6 we show the difference in F-measure of the SRL based on lexical data only (light grey bar) compared to the SRL using lexical and corpus data combined. The classes are ordered by decreasing difference in F-measure, so the difference is biggest for the leftmost class. The figure shows again that the improvement was most important for instances with very polysemous verbs (in many FCA classes). The gain in F-measure was lowest (rightmost bars in Figure B.6) for verbs in fewer classes but also for verbs in 23 or 25 classes, for these classes the labeling produced based on the FCA classification only had a better F-measure than when using the combined method. Overall, the number of FCA classes a verb is in does not seem to correlate with the SRL performance. In fact, the number of FCA concepts a verb is a member of may not necessarily be an indicator for the polysemy of this verb: due to the way FCA lattices are built, concepts often differ by only one subcategorisation frame. Therefore a verb may be in many classes and have few subcategorisation frames (and would be "less" polysemous) and also be in few classes

(a) SRL using frame and argument matcher

(b) SRL using frame and argument matcher combined with corpus frequency information.

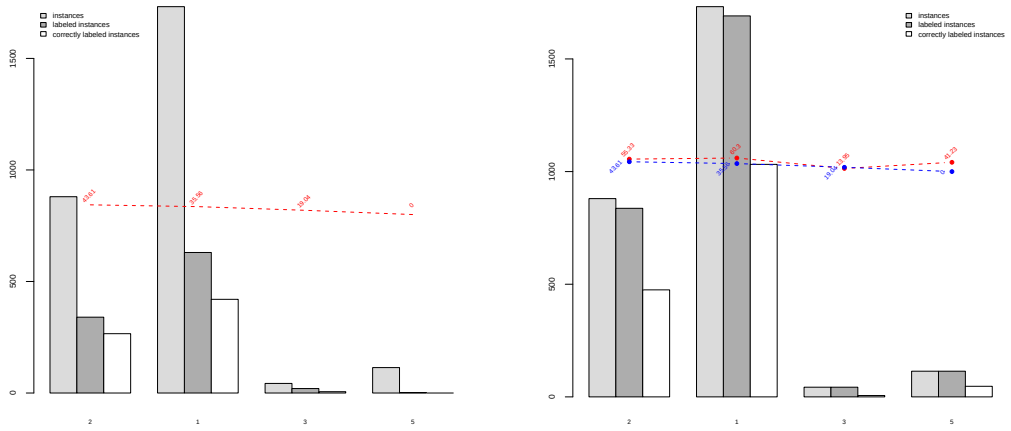


Figure B.7: Performance of SRL by verb translation polysemy classes (number of translated classes the verb is a member of).

but have many subcategorisation frames (and thus be "more" polysemous).

In the following discussion the polysemy class of a verb is the number of translated Verbnet classes it is a member of. The annotated verbs in the reference are members in 1, 2, 3 or 5 translated classes. Figure B.7 shows the number of annotated instances against these classes. More specifically, as before, the figure gives for each class (i) the number of annotated instances (light grey bars), (ii) the number of instances which could be labeled by the SRL system (darker grey bar), (iii) the number of correctly labeled instances (white bar) and (iv) the f-measure (red line and figures).

We observe that the polysemy of verbs with regard to the translated classes is lower than the polysemy w.r.t. the FCA classification. This is not surprising because in the FCA classification the syntactic frames are also taken into account: a verb may be a member of an FCA class only combined with a specific syntactic frame, which we did not consider in this analysis. Overall the f-measure is lower than for the best performing FCA polysemy classes, suggesting that FCA provides a more coherent verb, thematic role set association than the translation of Verbnet classes. However, the translated class polysemy does not seem to give further insights w.r.t. the accuracy of the labeling.

Bibliography

[Abeille *et al.*, 2003] Anne Abeille, Lionel Clément, and François Toussenel. *Building a treebank for French.* Kluwer, Dordrecht, 2003.

[Attik *et al.*, 2006] Mohammed. Attik, Shadi Al Shehabi, and Jean-Charles Lamirel. Clustering Quality Measures for Data Samples with Multiple Labels. In *Databases and Applications*, pages 58–65, 2006.

[Baker *et al.*, 1998] Collin F. Baker, Charles J. Fillmore, and John B. Lowe. The berkeley FrameNet project. In *Proceedings of the 17th International Conference on Computational Linguistics*, volume 1, pages 86–90, Montreal, Quebec, Canada, 1998. Association for Computational Linguistics.

[Barbut and Monjardet, 1970] Marc Barbut and Bernard Monjardet. *Ordre et Classification.* Hachette Université, 1970.

[Blanche-Benveniste *et al.*, 1984] Claire Blanche-Benveniste, Jos'e Delofeu, Jan Stefanini, and Karel van den Eynde. Pronom et syntaxe. l'approche pronominale et son application au fran cais. *SELAF*, 1984.

[Boons *et al.*, 1976] Jean-Paul Boons, Alain Guillet, and Christian Leclère. *La structure des phrases simples en français. I : Constructions intransitives.* Droz, Genève, 1976.

[Brew and Schulte im Walde, 2002] Chris Brew and Sabine Schulte im Walde. Spectral Clustering for German Verbs. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 117–124, Philadelphia, PA, 2002.

[Candito *et al.*, 2009] Marie Candito, Benoît Crabbé, and Mathieu Falco. Dépendances syntaxiques de surface pour le français. Technical report, Université de Paris 7, 2009.

[Chang and Lin, 2011] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2:27:1–27:27, 2011. Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

[Clear, 1993] Jeremy H. Clear. *The British national corpus*, pages 163–187. MIT Press, Cambridge, MA, USA, 1993.

[Dang *et al.*, 1998] Hoa Trang Dang, Karin Kipper, Martha Palmer, and Joseph Rosenzweig. Investigating Regular Sense Extensions Based on Intersective Levin Classes. In *COLING-ACL*, pages 293–299, 1998.

[Dempster *et al.*, 1977] Arthur P. Dempster, Nan M. Laird, and Donald. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, 39(1):1–38, 1977.

[Dorr, 1998] Bonnie J. Dorr. Large-scale dictionary construction for foreignlanguage tutoring and interlingual machine translation. *Machine Translation*, 12:271–322, January 1998.

[Dowty, 1991] David Dowty. Thematic proto-roles and argument selection. *Language*, 67:547–619, 1991.

[Eynde and Blanche-Benveniste, 1978] Karel Van den Eynde and Claire Blanche-Benveniste. Syntaxe et mécanismes descriptifs: présentation de l'approche pronominale. *Cahiers de Lexicologie*, 32(3–27), 1978.

[Falk and Gardent, 2010] Ingrid Falk and Claire Gardent. Bootstrapping a Classification of French Verbs Using Formal Concept Analysis. In *Interdisciplinary Workshop on Verbs Interdisciplinary Workshop on Verbs*, page 6, Pisa Italy, 11 2010.

[Falk and Gardent, 2011] Ingrid Falk and Claire Gardent. Combining Formal Concept Analysis and Translation to Assign Frames and Thematic Grids to French Verbs. In *Concept Lattices and their Applications*, Nancy, France, 10 2011.

[Falk *et al.*, 2010] Ingrid Falk, Claire Gardent, and Alejandra Lorenzo. Using Formal Concept Analysis to Acquire Knowledge about Verbs. In *Concept Lattices and their applications 7th International Conference on Concept Lattices and Their Applications - CLA 2010*, page 12, Sevilla, Spain, 10 2010.

[Falk *et al.*, 2012] Ingrid Falk, Claire Gardent, and Jean-Charles Lamirel. Classifying French Verbs Using French and English Lexical Resources. In *Proceedings of the 50th annual meeting of the Association for Computational Linguistics (ACL'12)*, Jeju, Republic of Corea, July 2012. The Association for Computational Linguistics (ACL).

[Fellbaum, 1998] Christiane Fellbaum, editor. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA, 1998.

[Fernandez *et al.*, 2002] Ana Fernandez, Gloria Vazquez, Patrick Saint-Dizier, Farah Benamara, and Mouna Kamel. The volem project: a framework for the construction of advanced multilingual lexicons. In *LEC '02: Proceedings of the Language Engineering Conference (LEC'02)*, page 89, Washington, DC, USA, 2002. IEEE Computer Society.

[Ferrer, 2004] Eva Esteve Ferrer. Towards a semantic classification of spanish verbs based on subcategorisation information. In *Proceedings of the ACL 2004 workshop on Student research*, ACLstudent '04, Stroudsburg, PA, USA, 2004. Association for Computational Linguistics.

[Fillmore *et al.*, 2003] Charles J. Fillmore, Christopher R. Johnson, and Miriam R. L. Petruck. Background to Framenet. *International Journal of Lexicography*, 16(3):235–250, 2003.

[Fritzke, 1995] Bernd Fritzke. A growing neural gas network learns topologies. *Advances in Neural Information Processing Systems 7*, 7:625–632, 1995.

[Fürstenau and Lapata, 2009] Hagen Fürstenau and Mirella Lapata. Graph Alignment for Semi-Supervised Semantic Role Labeling. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 11–20, Singapore, August 2009. Association for Computational Linguistics.

[Ganter and Wille, 1999] Bernhard Ganter and Rudolph Wille. *Formal concept analysis: Mathematical foundations*. Springer, Berlin-Heidelberg, 1999.

[Ghribi *et al.*, 2010] M. Ghribi, P. Cuxac, J.-C. Lamirel, and A. Lelu. Mesures de qualité de clustering de documents : prise en compte de la distribution des mots clés. In Nicolas Béchet, editor, *Évaluation des méthodes d'Extraction de Connaissances dans les Données- EvalECD'2010*, pages 15–28, Hammamet, Tunisie, January 2010. Fatiha Saïs.

[Gildea and Jurafsky, 2002] Daniel Gildea and Dan Jurafsky. Automatic labelling of semantic roles. *Computational Linguistics*, 2002.

[Giuglea and Moschitti, 2006] Ana-Maria Giuglea and Alessandro Moschitti. Semantic role labeling via FrameNet, VerbNet and PropBank. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, ACL-44, pages 929–936, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

[Gross, 1975] Maurice Gross. *Méthodes en syntaxe*. Hermann, Paris, 1975.

[Guillet and Leclère, 1992] Alain Guillet and Christian Leclère. *La structure des phrases simples en français. 2 : Constructions transitives locatives*. Droz, Geneva, 1992.

[Hebb, 1949] Donald O. Hebb. *The organization of behavior: a neuropsychological theory*. John Wiley & Sons, New York, 1949.

[Jackendoff, 1990] Ray S. Jackendoff. *Semantic Structures*. Cambridge: MIT Press, 1990.

[Joanis *et al.*, 2008] Eric Joanis, Suzanne Stevenson, and David James. A general feature space for automatic verb classification. *Natural Language Engineering*, 14(3):337–367, 2008.

[Kipper *et al.*, 2008] Karin Kipper, Anna Korhonen, Neville Ryant, and Martha Palmer. A large-scale classification of english verbs. *Language Resources and Evaluation*, 42(1):21–40, 2008.

[Klavans and Kan, 1998] Judith Klavans and Min-Yen Kan. Role of verbs in document analysis. In *Proceedings of the 17th international conference on Computational linguistics - Volume 1*, COLING '98, pages 680–686, Stroudsburg, PA, USA, 1998. Association for Computational Linguistics.

[Klimushkin *et al.*, 2010] Mikhail Klimushkin, Sergei Obiedkov, and Camille Roth. Approaches to the selection of relevant concepts in the case of noisy data. In Léonard Kwuida and Baris Sertkaya, editors, *Formal Concept Analysis*, volume 5986 of *Lecture Notes in Computer Science*, chapter 18, pages 255–266. Springer Berlin / Heidelberg, Berlin, Heidelberg, 2010.

[Korhonen *et al.*, 2003] Anna Korhonen, Yuval Krymolowski, and Zvika Marx. Clustering polysemic subcategorization frame distributions semantically. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics -*

Volume 1, ACL '03, pages 64–71, Stroudsburg, PA, USA, 2003. Association for Computational Linguistics.

[Korhonen, 2001] Anna Korhonen. *Subcategorization acquisition*. PhD thesis, University of Cambridge, Computer Laboratory, 2001.

[Korhonen, 2009] Anna Korhonen. Automatic Lexical Classification - Balancing between Machine Learning and Linguistics. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation*, Hong Kong, 2009.

[Kupść and Abeillé, 2008] Anna Kupść and Anne Abeillé. Growing treelex. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, volume 4919 of *Lecture Notes in Computer Science*, pages 28–39. Springer Berlin / Heidelberg, 2008.

[Kuznetsov, 2007] Sergei O. Kuznetsov. On stability of a formal concept. *Annals of Mathematics and Artificial Intelligence*, 49(1-4):101–115, 2007.

[Lamirel *et al.*, 2004] Jean-Charles Lamirel, Shadi Al Shehabi, Claire François, and Martial Hoffmann. New classification quality estimators for analysis of documentary information: application to patent analysis and web mapping. *Scientometrics*, 60(3):445–462, 2004. Article dans revue scientifique avec comité de lecture. internationale. A04-R-011 || lamirel04a.

[Lamirel *et al.*, 2008] Jean-Charles Lamirel, Anh Phuong Ta, and Mohammed Attik. Novel Labeling Strategies for Hierarchical Representation of Multidimensional Data Analysis Results. In *AIA - IASTED*, Innbruck, Autriche, 2008.

[Lamirel *et al.*, 2011a] Jean-Charles Lamirel, Pascal Cuxac, and Raghvendra Mall. A new efficient and unbiased approach for clustering quality evaluation. In *QIMIE'11, PaKDD*, Shenzen, China, 2011.

[Lamirel *et al.*, 2011b] Jean-Charles Lamirel, Raghvendra Mall, Pascal Cuxac, and Ghada Safi. Variations to incremental growing neural gas algorithm based on label maximization. In *Neural Networks (IJCNN), The 2011 International Joint Conference on*, pages 956 –965, 2011.

[Lang and Lapata, 2010] Joel Lang and Mirella Lapata. Unsupervised Induction of Semantic Roles. In *HLT-NAACL*, pages 939–947, 2010.

[Lang and Lapata, 2011a] Joel Lang and Mirella Lapata. Unsupervised Semantic Role Induction via Split-Merge Clustering. In *ACL*, pages 1117–1126, 2011.

[Lang and Lapata, 2011b] Joel Lang and Mirella Lapata. Unsupervised Semantic Role Induction with Graph Partitioning. In *EMNLP*, pages 1320–1331, 2011.

[Levin, 1993] Beth Levin. *English Verb Classes and Alternations: a preliminary investigation*. University of Chicago Press, Chicago and London, 1993.

[Li and Brew, 2008] Jianguo Li and Chris Brew. Which are the best features for automatic verb classification. In *ACL*, pages 434–442, 2008.

[Li, 2008] Jianguo Li. *Hybrid Methods for the Acquisition of Lexical Information: The Case for Verbs*. PhD thesis, Ohio State University, 2008.

[Loper *et al.*, 2007] Edward Loper, Szu ting Yi, and Martha Palmer. Combining lexical resources: Mapping between propbank and verbnet. In *Proceedings of the 7th International Workshop on Computational Linguistics*, 2007.

[Marcus *et al.*, 1993] Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. Building a large annotated corpus of english: the penn treebank. *Comput. Linguist.*, 19:313–330, June 1993.

[Martinetz and Schulten, 1991] Thomas Martinetz and Klaus Schulten. A "Neural-Gas" Network Learns Topologies. *Artificial Neural Networks*, I:397–402, 1991.

[Merlo and Stevenson, 2001] Paola Merlo and Suzanne Stevenson. Automatic verb classification based on statistical distributions of argument structure. *Comput. Linguist.*, 27:373–408, September 2001.

[Merlo and Van Der Plas, 2009] Paola Merlo and Lonneke Van Der Plas. Abstraction and generalisation in semantic role labels: PropBank, VerbNet or both? In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1 - Volume 1*, ACL '09, pages 288–296, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[Merlo *et al.*, 2002] Paola Merlo, Suzanne Stevenson, Vivian Tsang, and Gianluca Allaria. A Multilingual Paradigm for Automatic Verb Classification. In *ACL*, pages 207–214, 2002.

[Mertens, 2010] Piet Mertens. Restrictions de sélection et réalisations syntagmatiques dans DICOVALENCE. Conversion vers un format utilisable en TAL. . In *Actes TALN*, Montréal, 7 2010.

[Messiant *et al.*, 2008] Cédric Messiant, Anna Korhonen, and Thierry Poibeau. LexSchem: A Large Subcategorization Lexicon for French Verbs. In *Proceedings of the Language Resources and Evaluation Conference (LREC)*, Marrakech, 2008.

[Mouton, 2010] Claire Mouton. *Ressources et méthodes semi-supervisées pour l'analyse sémantique de textes en français*. PhD thesis, Université Pari 11 - Paris Sud UFR d'informatique, 2010.

[Oishi and Matsumoto, 1997] Akira Oishi and Yuji Matsumoto. Detecting the organization of semantic subclasses of Japanese verbs. *International Journal of Corpus Linguistics*, 2(1):65–89, october 1997.

[Padó and Lapata, 2009] Sebastian Padó and Mirella Lapata. Cross-lingual Annotation Projection for Semantic Roles. *J. Artif. Intell. Res. (JAIR)*, 36:307–340, 2009.

[Pado and Pitel, 2007] Sebastian Pado and Guillaume Pitel. Annotation précise du français en sémantique de rôles par projection cross-linguistique. In *Proceedings of TALN-07*, Toulouse, France, 2007.

[Pado, 2007] Sebastian Pado. *Cross-Lingual Annotation Projection Models for Role-Semantic Information*. PhD thesis, Saarland University, 2007.

[Palmer *et al.*, 2005] Martha Palmer, Paul Kingsbury, and Daniel Gildea. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106, 2005.

[Pereira *et al.*, 1993] Fernando Pereira, Naftali Tishby, and Lillian Lee. Distributional clustering of English words. In *Proceedings of the 31st annual meeting on Association for Computational Linguistics*, ACL '93, pages 183–190, Stroudsburg, PA, USA, 1993. Association for Computational Linguistics.

[Prudent and Ennaji, 2005] Yann Prudent and Abdellatif Ennaji. An incremental growing neural gas learns topologies. In *Neural Networks, 2005. IJCNN '05. Proceedings. 2005 IEEE International Joint Conference on*, volume 2, pages 1211–1216, 2005.

[Pustejovsky, 1995] James Pustejovsky. *The Generative Lexicon*. The Mit Press, 1995.

[Randall, 2010] Janet H. Randall. *Linking*. Studies in Natural Language and Linguistic Theory. Springer, Dordrecht, 2010.

[Robertson and Sparck Jones, 1976] Steven E. Robertson and Karen Sparck Jones. Relevance weighting of search terms. *Journal of the American Society for Information Science*, 27(3):129–146, 1976.

[Rooth *et al.*, 1999] Mats Rooth, Stefan Riezler, Detlef Prescher, Glenn Carroll, and Franz Beil. Inducing a semantically annotated lexicon via em-based clustering. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 104–111, Stroudsburg, PA, USA, 1999. Association for Computational Linguistics.

[Roth *et al.*, 2006] Camille Roth, Sergei A. Obiedkov, and Derrick G. Kourie. Towards concise representation for taxonomies of epistemic communities. In *CLA*, pages 240–255, 2006.

[Sagot *et al.*, 2006] Benoît Sagot, Lionel Clément, Éric Villemonte de La Clergerie, and Pierre Boullier. The Lefff 2 syntactic lexicon for French: architecture, acquisition, use. In *Proc. of LREC'06*, 2006.

[Saint-Dizier, 1996] Patrick Saint-Dizier. Constructing verb semantic classes for french: Methods and evaluation. In *COLING*, pages 1127–1130, 1996.

[Saint-Dizier, 1999] Patrick Saint-Dizier. Alternation and verb semantic classes for french: Analysis and class formation. In *Predicative forms in natural language and in lexical knowledge bases*. Kluwer Academic Publishers, 1999.

[Samardžić, 2009] Tanja Samardžić. Semantic roles in natural language processing. Master's thesis, Université de Genève, Departement de Linguistique, 10 2009.

[Schuler, 2006] Karin Kipper Schuler. *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon*. PhD thesis, University of Pennsylvania, 2006.

[Schulte im Walde *et al.*, 2008] Sabine Schulte im Walde, Christian Hying, Christian Scheible, and Helmut Schmid. Combining em training and the mdl principle for an automatic verb classification incorporating selectional preferences. In *Proceedings of the 46th Annual Meeting of the ACL: Human Language Technology*, Columbus, Ohio, 2008.

[Schulte im Walde, 2003] Sabine Schulte im Walde. *Experiments on the Automatic Induction of German Semantic Verb Classes*. PhD thesis, Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart, 2003. Published as AIMS Report 9(2).

[Schulte im Walde, 2006] Sabine Schulte im Walde. Experiments on the automatic induction of german semantic verb classes. *Computational Linguistics*, 32(2):159–194, 2006.

[Schulte im Walde, 2009] Sabine Schulte im Walde. The Induction of Verb Frames and Verb Classes from Corpora. In Anke Lüdeling and Merja Kytö, editors, *Corpus Linguistics. An International Handbook.*, number 29 in Handbooks of Linguistics and Communication Science, chapter 44, pages 952–971. Mouton de Gruyter, Berlin, March 2009.

[Snider and Diab, 2006] Neal Snider and Mona Diab. Unsupervised induction of modern standard arabic verb classes. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, NAACL-Short '06, pages 153–156, Stroudsburg, PA, USA, 2006. Association for Computational Linguistics.

[Sun and Korhonen, 2009] Lin Sun and Anna Korhonen. Improving verb clustering with automatically acquired selectional preferences. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2*, EMNLP '09, pages 638–647, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.

[Sun and Korhonen, 2011] Lin Sun and Anna Korhonen. Hierarchical verb clustering using graph factorization. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 1023–1033, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[Sun *et al.*, 2008] Lin Sun, Anna Korhonen, and Yuval Krymolowski. Verb class discovery from rich syntactic data. In *Proceedings of the 9th international conference on Computational linguistics and intelligent text processing*, CICLing'08, pages 16–27, Berlin, Heidelberg, 2008. Springer-Verlag.

[Sun *et al.*, 2010] Lin Sun, Anna Korhonen, Thierry Poibeau, and Cédric Messiant. Investigating the cross-linguistic potential of VerbNet-style classification. In *Proceedings of the 23rd International Conference on Computational Linguistics*, COLING '10, pages 1056–1064, Stroudsburg, PA, USA, 2010. Association for Computational Linguistics.

[Swier and Stevenson, 2004] Robert S. Swier and Suzanne Stevenson. Unsupervised Semantic Role Labeling. In *EMNLP*, pages 95–102, 2004.

[Swier and Stevenson, 2005] Robert S. Swier and Suzanne Stevenson. Exploiting a verb lexicon in automatic semantic role labelling. In *HLT/EMNLP*, 2005.

[Swift, 2005] Mary Swift. Towards automatic verb acquisition from VerbNet for spoken dialog processing. In *Proceedings of Interdisciplinary Workshop on the Identification and Representation of Verb Features and Verb Classes*, pages 115–120, Saarbrücken, 2005.

[Titov and Handerson, 2010] Ivan Titov and James Handerson. A Latent Variable Model for Generative Dependency Parsing. In H. Bunt, P. Merlo, and J. Nivre, editors, *Trends in Parsing Technology*. Text, Speech and Language Technology Series (Springer), 2010.

[van den Eynde and Mertens, 2003] Karel van den Eynde and Piet Mertens. La valence : l'approche pronominale et son application au lexique verbal. *Journal of French Language Studies*, 13:63–104, 2003.

[van der Plas *et al.*, 2011] Lonneke van der Plas, Paola Merlo, and James Henderson. Scaling up Automatic Cross-Lingual Semantic Role Annotation. In *ACL (Short Papers)*, pages 299–304, 2011.

[Vlachos *et al.*, 2009] Andreas Vlachos, Anna Korhonen, and Zoubin Ghahramani. Unsupervised and Constrained Dirichlet Process Mixture Models for Verb Clustering. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, GEMS '09, pages 74–82, Stroudsburg, PA, USA, 2009. Association for Computational Linguistics.