Modélisation probabiliste de l'interface syntaxe sémantique à l'aide de grammaires hors contexte probabilistes
Expériences avec FrameNet

Olivier Michalon
LIF, AMU, CNRS, 163 avenue de Luminy, 13288 Marseille Cedex 9
olivier.michalon@lif.univ-mrs.fr

Résumé. Cet article présente une méthode générative de prédiction de la structure sémantique en cadres d'une phrase à partir de sa structure syntaxique et décrit les grammaires utilisées ainsi que leurs performances. Ce système permet de prédire, pour un mot dans le contexte syntaxique d'une phrase, le cadre le plus probable. Le système génératif permet d'attribuer à ce mot un cadre et à l'ensemble de chemins des rôles sémantiques. Bien que les résultats ne soient pas encore satisfaisants, cet analyseur permet de regrouper les tâches d'analyse sémantique (sélection du cadre, sélection des actants, attribution des rôles), contrairement aux travaux précédemment publiés. De plus, il offre une nouvelle approche de l'analyse sémantique en cadres, dans la mesure où elle repose plus sur la structure syntaxique que sur les mots de la phrase.

Abstract. This paper presents a generative method for predicting the frame semantic structure of a sentence from its syntactic structure and describes the grammars used with their performances. This system allows to predict, for a word in the syntactic context of a sentence, the most probable frame. The generative system allows to give a frame to a word and semantic roles to a set of pathes. Although results are not yet satisfying, this parser allows to group semantic parsing tasks (frame selection, role fillers selection, role assignment) unlike previously published works. In addition, it offers a new approach to parse semantic frames insofar as it is based more on syntactic structure rather than words of the sentence.

Mots-clés : Analyse sémantique automatique, interface syntaxe sémantique, FrameNet.

Keywords: Automatic Semantic Parsing, syntax semantic parsing, FrameNet.

1 Introduction

Plusieurs théories de représentation de la structure sémantique coexistent actuellement, et parmi celles-ci, plusieurs projets ont vu le jour, parmi lesquels on trouve WordNet (Miller & Fellbaum, 1998), PropBank (Palmer *et al.*, 2005) ou encore FrameNet (Fillmore & Baker, 2001) . Alors que WordNet s'attache à représenter la hiérarchie entre différents noms et verbes (une voiture est une sous-classe de véhicule à moteur), PropBank associe une étiquette à certains actants syntaxiques des verbes. Le projet FrameNet quant à lui se base sur une analyse des situations nommées cadres sémantiques. Ces cadres sémantiques représentent chacun une action ou un concept, en incluant les éléments qui y jouent un rôle. Par exemple pour l'action *manger*, le cadre sémantique identifie la *personne qui mange*, la *nourriture*, ainsi que les *outils utilisés pour manger*. Ces différents rôles ne sont pas sans rapport avec les arguments de ProbBank mais l'inventaire est beaucoup plus riche et FrameNet recense des relations entre les rôles de différents cadres.
Chacun de ces projets a permis de créer des données annotées qui permettent d'estimer les paramètres d'outils d'analyse automatique basés sur les données.

L'approche que nous présentons dans cet article consiste à utiliser les données FrameNet de l'anglais (le corpus Français est en cours de construction) pour comprendre à quel point la structure syntaxique d'une phrase peut jouer un rôle dans la prédiction de sa structure sémantique. En effet la réalisation syntaxique d'actants sémantiques présente des régularités et ce sont précisément ces régularités que nous cherchons à modéliser ici. Pour l'action *manger*, la personne qui mange sera la plupart du temps le sujet du verbe, la nourriture sera l'objet et les outils seront des compléments. Pour automatiser cette tâche, nous utilisons les données annotées du corpus FrameNet pour estimer les probabilités d'une grammaire probabiliste hors contexte. Cette grammaire permet d'analyser des réalisations syntaxiques en leur associant la représentation sémantique la plus probable. Contrairement aux travaux réalisés par Modi *et al.* (2012), ou encore Das *et al.* (2010), dans lesquels la tâche d'analyse sémantique est scindée en trois voire en quatre parties distinctes (identification des mots dé-

clencheurs de cadre, sélection du cadre, identifications des mots acteurs du cadre, attribution des rôles à ces mots), notre grammaire permet de réaliser ces opérations simultanément, ce qui permet d'envisager l'analyse sémantique comme une tâche globale.
Nous ne nous attacherons dans cette étude qu'à trois parties de discours auxquelles on peut attribuer des cadres sémantiques : les adjectifs (*JJ*), les noms (*NN*), et les verbes (*VS*). Nous avons fait ce choix car ces trois parties de discours sont celles présentant la plus grande variété sémantique et syntaxique.

Après avoir présenté la théorie et les données de FrameNet, nous détaillerons le modèle que nous avons utilisé pour réaliser l'analyse sémantique. Nous continuerons avec les résultats de ce modèle et conclurons en proposant quelques pistes pour la suite.

2 FrameNet

2.1 Présentation du projet FrameNet

Le projet FrameNet regroupe à la fois une théorie de modélisation de la sémantique des phrases et un ensemble de données annotées manuellement. L'unité structurelle utilisée pour représenter la structure sémantique d'une phrase est appelée *Semantic Frame* (cadre sémantique en français), et elle est fondée sur les travaux de Charles J. Fillmore (Fillmore, 1976, 1977, 1982, 1985; Fillmore & Baker, 2010).

Par exemple, le verbe *continuer* (*continue*) peut évoquer deux situations différentes : *Une activité qui continue* (modélisée par le cadre *Process_continue*) ou *la poursuite d'une activité* (*Activity_ongoing*). Ces situations, bien que proches, ne représentent pas la même chose : la première concerne une tâche qui se poursuit, alors que la seconde concerne des participants qui continuent une activité.
Le cadre sémantique de *Process_continue* met en jeu essentiellement un *événement* (*Event*), mais peut aussi contenir des *circonstances* (*Circumstances*), une *manière de continuer* (*Manner*), ou encore *l'événement suivant* (*Next_subevent*), comme dans l'exemple suivant :

$$\text{The meeting}_{Event}\ \text{continued}_{PROCESS\_CONTINUE}\ \text{with a discussion}_{Next\_subevent}.$$
(La réunion s'est poursuivie par un débat.)

Le cadre sémantique de *Activity_ongoing* nécessite absolument de définir *l'activité* (*Activity*) et un *agent* (*Agent*) effectuant cette activité. Bien entendu ce cadre sémantique peut aussi contenir des indicateurs de *circonstances* (*Circumstances*), d'une *manière de continuer* (*Manner*), ou encore de *l'événement suivant* (*Next_subevent*), comme dans l'exemple suivant :

$$\text{We}_{Agent}\ \text{continued}_{ACTIVITY\_ONGOING}\ \text{the meeting}_{Event}\ \text{with a discussion}_{Next\_subevent}.$$
(Nous avons poursuivi la réunion par un débat.)

Les deux situations évoquées précédemment (*Process_continue* et *Activity_ongoing*) sont ce que nous appellerons des *cadres* (*frame* en anglais) et les intervenants (*Event*, *Agent*,...) sont appelés *rôles* (*frame elements*). Les mots qui pourront évoquer ces cadres (comme *poursuivre*) sont appelés *ancres* (*lexical units*) du cadre. Les rôles sont spécifiques à un cadre. Le mot ou l'expression qui joue un rôle dans un cadre est appelé *acteur*. Le nombre de rôles peut varier selon les cadres. Certains rôles peuvent ne pas être réalisés pour une occurrence de cadre donnée, le jeu de rôles instanciés peutdonc varier d'une occurrence de cadre à une autre. Lorsqu'une ancre est associée à un cadre, on dit que l'ancre déclenche le cadre, qui est alors instancié.
Notons également qu'un mot ne peut déclencher qu'un seul cadre alors qu'un mot peut occuper un rôle dans plusieurs cadres. De plus, un mot peut être à la fois ancre et acteur.

Le projet FrameNet a permis de créer des ressources lexicales et un corpus annoté pour l'anglais. En ce qui concerne la version 1.5 du projet, les données sont composées :

- d'un lexique associant une ancre à des cadres ;
- d'un inventaire de cadres et de leurs rôles ;
- d'un corpus annoté (*corpus FullText*) en cadres et étiqueté en parties de discours ;
- d'un ensemble de données d'exemple annotées pour des cadres spécifiques.

Le *lexique* comprend un peu moins de 12000 entrées. Chacune des entrées associe un cadre à une unité lexicale (un lemme avec une partie de discours). Une unité lexicale peut être associée à plusieurs cadres, comme les exemples précédents nous l'ont montré. On trouve en fin de compte 894 cadres différents et 9394 ancres différentes. Le tableau 1 regroupe trois informations importantes pour chaque partie de discours que nous allons traiter : le nombre d'ancres, le nombre de cadres

	Nombre d'entrées	Nombre de cadres	Ambiguïté moyenne
Adjectifs	1850	307	1.15
Verbes	3060	604	1.5
Noms	4166	605	1.14

TABLE 1 – Variété des ancres selon leur partie de discours dans la version 1.5 de FrameNet

pouvant être déclenchés, et l'ambiguïté moyenne, qui est le nombre moyen de cadres pouvant être déclenchés par ancre.
L'*inventaire* de cadres sémantique recense chaque cadre, liste et définit les rôles lui appartenant ainsi que les liens existants avec d'autres cadres.
Le *corpus* est composé d'un ensemble de phrases dont la structure sémantique est annotée manuellement. On y trouve aussi une annotation en parties de discours des phrases. Ce corpus est composé de 4037 phrases. On y trouve 23871 occurrences de cadres (706 cadres différents), et 46929 occurrences d'ancres (3345 ancres différentes). Le fait que le nombre d'occurrences d'ancres soit bien supérieur au nombre d'instances de cadres provient du fait que les ancres peuvent parfois véhiculer un sens qui n'est pas modélisé dans la théorie FrameNet.
Le *corpus d'exemples* est un corpus annoté à la main pour des cadres spécifiques. Chaque phrase de ce corpus est destinée à illustrer l'utilisation d'un seul cadre, par conséquent seul un cadre sera annoté par phrase. Ce corpus dispose lui aussi d'annotations en parties de discours.

2.2 Prétraitements effectués sur le corpus

Pour nos travaux nous utilisons le corpus de phrases annotées complètement en lui ajoutant une analyse syntaxique et les lemmes, puis nous le scinderons en un corpus d'entraînement et un corpus de test.

Du corpus FrameNet, nous utiliserons les mots, parties de discours, ainsi que les cadres sémantiques. Pour ce qui est des parties de discours, nous les normalisons pour qu'elles correspondent toutes à celles du *PennTreeBank*.
Pour savoir si un mot peut être une ancre ou non, nous avons besoin de connaître sa forme lemmatisée, nous ajoutons donc les lemmes des mots aux données extraites du corpus FrameNet. Notre approche s'appuyant sur la syntaxe, nous effectuons également une analyse syntaxique en dépendances de notre corpus. La lemmatisation et l'analyse en dépendances ont été réalisées avec le logiciel *Mate-tools* (Bohnet *et al.*, 2013; Bohnet, 2010), dont la performance en analyse syntaxique avec étiquettes de dépendances est à $90,33\%$ sur le corpus *Conll Shared Task 2009*. Nous devrons donc composer avec les erreurs syntaxiques générées par l'analyseur. Notons également que les lemmes de ce corpus sont automatiquement prédits.

Dans les annotations FrameNet, un acteur est souvent associé à un segment de phrase, généralement un syntagme. Par souci de simplicité, les acteurs composés de plusieurs mots ne seront plus que représentés par leur tête syntaxique. Bien que cette réduction puisse être vue comme une perte d'information, dans la plupart des cas la tête syntaxique correspond bien à la tête sémantique de l'expression.

Le corpus est ensuite séparé en deux : (*corpus d'entraînement* : 3055 phrases et *corpus de test* : 982 phrases) selon la séparation historique (Das & Smith, 2011) établie pour la tâche SemEval. On extrait ensuite du corpus d'entraînement l'ensemble des chemins syntaxiques (*corpus des chemins*) permettant d'aller d'une ancre aux acteurs correspondants.

Un *chemin syntaxique* correspond au chemin emprunté pour aller d'un noeud à un autre dans l'arbre de dépendances syntaxiques de la phrase. Il est possible d'aller dans le sens d'une dépendance syntaxique ou d'aller à contre courant (dépendant → gouverneur). Chaque branche de l'arbre peut être empruntée 0 ou 1 fois, ce qui garantit l'unicité du chemin entre deux noeuds donnés. Si une dépendance est empruntée à contre courant, on l'indique en faisant précéder l'étiquette de dépendance syntaxique du signe « - ». Un chemin syntaxique est donc défini comme une séquence *[(+/-), lex, lemme, PoS, fct]**, où le symbole de signe indique le sens de la dépendance, *lex* (et *lemme*) le mot (et son lemme) traversé par ce chemin, *PoS* sa partie de discours et *fct* sa fonction syntaxique. Pour nos expériences nous souhaitons faire des statistiques sur les chemins syntaxiques. Le corpus étant de taille réduite et afin d'obtenir des statistiques fiables, nous ne pouvons nous permettre d'avoir des chemins d'une grande variété mais avec une fréquence faible. Pour cela nous allons les simplifier afin de faire des regroupements plus conséquents. Nous réduirons donc les chemins syntaxiques à des suites d'étiquettes de dépendances syntaxiques orientées.

Par exemple, dans la phrase suivante :

Maurice$_{Cook}$ bakes$_{APPLY_HEAT}$ an apple pie$_{Food}$ in the oven$_{Heating_instrument}$.
(Maurice cuit une tarte aux pommes dans le four.)

Le mot *bakes* est une ancre pour le cadre *Apply_heat*, dont les rôles instanciés ici sont *Cook*, *Food* et *Heating_instrument*, dont les acteurs sont respectivement *Maurice*, *apple pie* (dont la tête syntaxique est *pie*) et *oven*.
L'arbre syntaxique de la phrase est représenté en figure 1, et les chemins associés au cadre *Apply_heat* sont décrits dans la table 2.

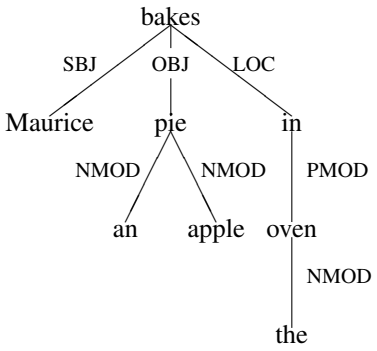

FIGURE 1 – Arbre syntaxique de la phrase *Maurice bakes an apple pie in the oven*

Rôle	**Début du chemin**	→	**Fin du chemin (Acteur)**	**Représentation du chemin**
Cook	bakes	→	Maurice	(+SBJ)
Food	bakes	→	pie	(+OBJ)
Heating_instrument	bakes	→	oven	(+LOC,+PMOD)

TABLE 2 – Chemins syntaxiques du cadre *Apply_heat* pour la phrase *Maurice bakes an apple pie in the oven.*

La table 3 présente les cinq chemins les plus fréquents pour les ancres des parties de discours que nous traitons (*Noms*, *Adjectifs* et *Verbes*). On remarque tout d'abord que certains chemins syntaxiques sont effectivement prédominants par rapport aux autres. Pour les noms et les adjectifs, on remarque également que le chemin vide (représenté *()*) est très représenté : les ancres ayant ces parties de discours sont donc souvent des acteurs des cadres qu'ils déclenchent. On remarque aussi que l'épithète du nom (*NMOD*) lorsque l'ancre est un nom joue souvent un rôle. De même lorsque l'ancre est un adjectif, le nom que qualifie l'adjectif est souvent acteur (*-NMOD*). En ce qui concerne les verbes, les traditionnels sujet (*SBJ*) et objet (*OBJ*) sont les éléments qui jouent le plus souvent un rôle.

Noms		Adjectifs		Verbes	
(NMOD)	5702/13381	(-NMOD)	1418/3366	(OBJ)	2696/12679
()	5213/13381	()	921/3366	(SBJ)	1802/12679
(LOC)	170/13381	(AMOD)	206/3366	(-VC,SBJ)	968/12679
(-NMOD,NMOD)	161/13381	(-NMOD,NMOD)	122/3366	(ADV)	858/12679
(-OBJ,SBJ)	113/13381	(-PRD,SBJ)	119/3366	(OPRD)	555/12679

TABLE 3 – Les cinq chemins syntaxiques les plus fréquents dans le corpus d'entraînement en fonction de la partie de discours de l'ancre

Si, pour une instanciation de cadre sémantique, on regroupe tous les chemins syntaxiques des rôles de ce cadre, on obtient une représentation linéaire de la structure syntaxique de ce cadre (à condition de connaître l'origine des chemins syntaxiques de ce cadre, à savoir l'ancre). Nous appelons cette représentation **signature syntaxique** du cadre, en voici un exemple pour la phrase *« Maurice bakes an apple pie in the oven. »*.

[bakes, (+SBJ), (+OBJ), (+LOC,+PMOD)]

Remarquons que l'ordre d'apparition des différents chemins syntaxiques n'a pas d'importance ici. On dit que la signature est **centrée sur l'ancre**.

3 Modèle

Notre approche consiste à modéliser les régularités syntaxiques des structures sémantiques du corpus d'entraînement afin de prédire les structures sémantiques du corpus de test. Pour cela, nous allons dans un premier temps proposer un système de prédiction de cadres sémantiques qui attache un cadre à un mot et des rôles à certains mots de la phrase. Ce système utilisera les données syntaxiques pour faire ses choix. Dans un second temps, nous évaluerons les performances de ce système. Pour rappel, nous ne cherchons à déterminer que les cadres dont les ancres ont pour parties de discours *Adjectif*, *Nom* ou *Verbe*.

Pour chaque ancre a apparaissant dans le corpus de test nous évaluons un certain nombre de possibilités d'étiquetages sémantiques et sélectionnons celle à laquelle le modèle attribue le meilleur score. L'étiquetage sémantique consiste à associer un cadre à a et des chemins syntaxiques aux rôles de ce cadre. En termes plus mathématiques, nous allons déterminer la réalisation de cadre $\widehat{F}$ la plus probable compte tenu d'une signature syntaxique S centrée sur l'ancre a :

$$\widehat{F} = \underset{F}{argmax} P(F|S)$$

$\widehat{F}$ est calculé en énumérant puis comparant toutes les représentations sémantiques compatibles avec la signature S. Pour énumérer toutes les représentations sémantiques possibles, nous créons une grammaire hors contexte qui reconnaît toutes les signatures possibles (symboles terminaux) en leur associant une structure sémantique (symboles non terminaux). La figure 2 illustre une dérivation possible de cette grammaire. Cette approche générative a deux intérêts principaux : elle produira toutes les analyses possibles et sera facile à modifier. La facilité de modification permettra de tester plusieurs hypothèses d'indépendance et d'améliorer l'analyse, à l'instar de Collins et de ses modèles génératifs d'analyse syntaxique (Collins, 1997).

Notre grammaire peut générer toutes les signatures composées d'une ancre et d'une suite de chemins syntaxiques (de la forme *[ancre, chemin$_1$, chemin$_2$, ..., chemin$_n$]*). Cette grammaire a pour axiome le symbole $\mathcal{F}$ pouvant se réécrire en chaque cadre. Chaque cadre peut être réécrit en une ancre (parmi les ancres possibles du cadre) et en un ensemble de rôles spécifiques à ce cadre. Chaque rôle peut être réécrit en chemin syntaxique. Dans cette grammaire, les cadres et les rôles sont donc des symboles non terminaux. Une signature reconnue par cette grammaire peut être reconnue de plusieurs façons différentes, chacune correspondant à un étiquetage sémantique particulier.

Afin de comparer les différents étiquetages sémantiques d'un même mot, nous leur attribuons une probabilité. Nous assignons alors à chaque règle de la grammaire une probabilité. Notre grammaire devient alors une grammaire hors contexte probabiliste. Les probabilités assignées à chaque règle sont estimées à partir du corpus d'entraînement. Pour sélectionner le meilleur étiquetage sémantique, nous utilisons l'algorithme d'Earley (Earley, 1970), qui calcule simultanément toutes les analyses et les représente sous forme de forêt partagée en un temps polynomial ($O(n^3)$ avec n le nombre de symboles non terminaux composant la signature). La représentation en forêt partagée est utile à la fois pour trouver la meilleure analyse et pour connaître le score d'une analyse partielle (notamment le score d'un cadre toutes sous-catégorisations confondues). La figure 2 représente un des arbres de dérivation pour la phrase *Maurice bakes an apple pie in the oven*. On remarque que l'axiome se réécrit en *Apply_heat*, qui lui même se réécrit en *Apply_heat(Verbe)*, introduisant la partie de discours de l'ancre avant de l'avoir générée.

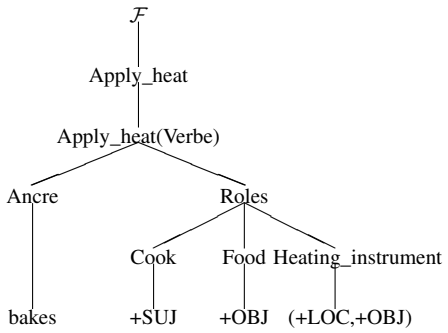

FIGURE 2 – Un arbre de dérivation de la phrase *Maurice bakes an apple pie in the oven* avec notre grammaire, en sélectionnant *bakes* comme ancre.

Notre grammaire permet maintenant de reconnaître des signatures syntaxiques en leur attribuant une structure sémantique. Cette grammaire permet d'assigner un cadre à l'ancre et un rôle à chaque chemin de la signature. Les rôles attribués sont soit des rôles appartenant au cadre, soit des rôles que nous qualifierons de vides si le chemin syntaxique correspondant ne participe pas à la structure sémantique.

Deux points importants restent à éclaircir, à savoir comment limiter le nombre de chemins et comment représenter le cas d'une ancre ne déclenchant pas de cadre.

Le cas de l'ancre ne déclenchant pas de cadre Rappelons que les ancres sont des mots répertoriés dans le lexique FrameNet et pouvant entraîner l'apparition d'un cadre sémantique. Certains occurrences de ces mots se font dans des situations dont la sémantique ne correspond pas à des cas définis sémantiquement. Ce cas est très présent (2070 occurrences pour 5504 ancres) et il faut donc modéliser dans nos grammaires ces cas. Notre approche est simple : nous allons créer un nouveau cadre, que nous nommerons le *cadre nul*. Ce cadre sera légèrement différent des autres cadres puisqu'il ne possèdera aucun rôle. Il ne possédera donc aucune réalisation syntaxique. Pour le prendre en compte, nous allons tout simplement estimer la probabilité qu'une ancre donnée déclenche le cadre nul à partir du corpus d'entraînement.

Sélection des chemins utilisables Pour sélectionner les chemins utilisables, nous ne prenons que les chemins liant une ancre à un rôle apparus au moins deux fois dans le corpus. Nous avons fait ce choix pour éliminer les chemins aberrants dûs à des erreurs d'analyse (l'analyseur en dépendance se trompe dans presque 1 cas sur 10 sur le corpus *Conll Shared Task 2009*) sans pour autant nous priver de données précieuses. En effet, le corpus d'entraînement comprend 2401 chemins ancre-rôle différents, et comme le montre le tableau 4, la variété des chemins décroît fortement lorsque leur nombre d'occurrences augmente. Ce choix permet aussi de limiter la taille des signatures que nous soumettons à l'algorithme d'Earley ainsi que la taille des grammaires qui sont utilisées.

occurrences	Nombre de chemins différents
> 0	2401
> 1	648
> 2	405
> 3	312
> 4	263
> 5	227

TABLE 4 – Nombre de chemins différents en fonction de leur nombre d'occurrences minimal

Exemple Prenons la phrase :

> We continued the meeting with a discussion

Si on se concentre sur l'ancre *continued*, il existe 7 chemins ayant cette ancre pour origine dans cette phrase. Admettons que parmi ces chemins seuls les suivants aient été observés plus d'une fois dans le corpus d'entraînement :

- $continued \rightarrow We$: +SBJ
- $continued \rightarrow meeting$: +OBJ
- $continued \rightarrow with$: +ADV
- $continued \rightarrow discussion$: +ADV,+PMOD

La signature que l'on soumettrait à la grammaire serait donc :

> continued,(+SBJ),(+OBJ),(+ADV),(+ADV,+PMOD)

L'algorithme d'Earley construit une forêt composée de tous les arbres possibles pour générer cette signature, la figure 3 présente deux des arbres de cette forêt d'analyses.

Les probabilités associées à chacun de ces arbres sont comparées afin de choisir la meilleure. Idéalement l'algorithme devrait choisir le cadre *activity_ongoing*, grâce à la présence d'un complément d'objet direct dans la phrase.

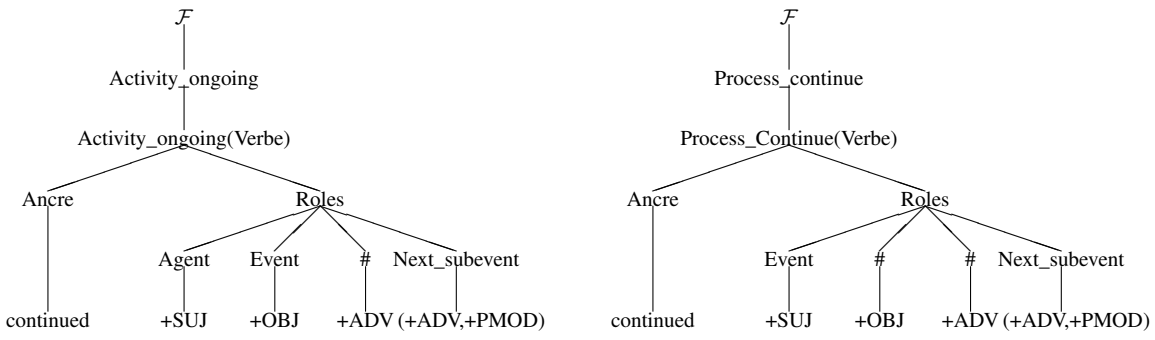

FIGURE 3 – Deux des arbres de la forêt d'analyses de la phrase *We continued the meeting with a discussion*. Le symbole # représente le rôle vide, c'est à dire les chemins qui ne correspondent à aucun rôle.

3.1 Grammaires utilisées

Pour nos expériences, nous avons créé trois grammaires différentes. La principale distinction entre ces grammaires tient dans la nature des interactions entre les rôles d'un même cadre. Nous appellerons *sous-catégorisation* l'ensemble des rôles observés pour une occurrence de cadre. La première grammaire va donner une importance capitale aux sous-catégorisations observées dans le corpus d'entraînement, alors que la seconde n'en tiendra absolument pas compte. La troisième grammaire sera plus nuancée, et plutôt que de s'attacher à la composition des sous-catégorisations, elle s'attachera au nombre de rôles réalisés dans chacune d'elles. Pour faire simple, nous avons deux grammaires extrêmes : la première qui suit scrupuleusement les sous-catégorisations observées lors de l'entraînement et la deuxième qui n'en tient aucunement compte. La troisième grammaire est en quelque sorte un compromis entre les deux. L'hypothèse commune à ces trois grammaires est que la réalisation syntaxique d'un rôle est indépendante de la réalisation des autres rôles instanciés pour le même cadre. Lorsqu'un rôle est attribué à une réalisation syntaxique, on ne regarde pas quels sont les autres réalisations syntaxiques sélectionnées. Une limite de cette approche provient du fait que seuls les cadres ayant été observés dans le corpus d'entraînement peuvent être générés par nos grammaires.

L'axiome de la grammaire est noté $\mathcal{F}$. Parmi les symboles non terminaux, les cadres seront notés F, les ancres A, les sous-catégorisations S, et les rôles R. Les symboles terminaux sont les mots du lexique et les chemins dans le format défini plus tôt, que nous désignerons respectivement par mot et $chemin$

Bien entendu les grammaires complètes ne sont pas présentées ici car trop vastes, les cadres sont alors numérotés de 1 à m et les chemins de 1 à p. Pour chaque cadre i, les ancres seront notées de 1 à a_i ,les rôles de 1 à r_i et les sous-catégorisations de 1 à s_i. Notons que la partie de discours de l'ancre influe sur les règles de réécriture des rôles en chemins.

3.1.1 G1 : Grammaire proche des sous-catégorisations

Cette grammaire cherche avant tout à reproduire les sous-catégorisations observées dans le corpus. Elle totalise 35468 règles.

	Probabilité	**Règle**	
1	$P(F_i)$	$\mathcal{F} \rightarrow F_i$	$\forall i \in 1,...,m$
2	$P(F_{nul})$	$\mathcal{F} \rightarrow F_{nul}$	
3	$P(F_{i,p}\|F_i)$	$F_i \rightarrow F_{i,p}$	$\forall i \in 1,...,m,\ p \in \{NN, JJ, VB\}$
4	$P(A_j, S_k\|F_{i,p}) = 1$	$F_{i,p} \rightarrow A_j S_k$	$\forall i \in 1,...,m,\ j \in 1,...,a_i,\ k \in 1,...,s_i$
5	$P(A_j\|F_{nul}) = 1$	$F_{nul} \rightarrow A_j$	$\forall j \in 1,...,a_{nul}$
6	$P(mot\|A_j)$	$A_j \rightarrow mot$	$\forall i \in 1,...,m,\ j \in 1,...,a_i$
7	$P(R_x, R_y, ...\|S_k)$	$S_k \rightarrow R_x R_y ...$	$\forall k \in 1,...,s_i,\ i \in 1,...,m;\ x, y \in [1, r_i], x \neq y$
8	$P(chemin\|R_r, p)$	$R_r \rightarrow chemin$	$\forall r \in 1,...,r_i,\ i \in 1,...,m$

Cette grammaire se base sur l'hypothèse que seules les sous-catégorisations observées lors de l'entraînement sont possibles. Pour que la grammaire accepte une signature, il faut que les chemins proposés correspondent chacun à un rôle

d'une sous-catégorisation observée dans le corpus d'entraînement, et que cette sous-catégorisation voie tous ses rôles instanciés.

Les probabilités des deux premières règles correspondent simplement à la probabilité d'un cadre d'apparaître. La règle numéro 3 permet de fixer la partie de discours de l'ancre de ce cadre pour la suite des règles, de façon à ce que les chemins soient cohérents avec cette partie de discours. La règle 4 permet de réécrire le cadre enrichi d'une partie de discours en deux symboles : l'un représentant les ancres possibles pour cette catégorie et pour ce cadre, l'autre représentant une sous-catégorisation possible pour ce cadre. L'ancre est réécrite sous la forme d'un lemme (symbole terminal), et la sous-catégorisation se réécrit en un ensemble de rôles. Chaque rôle se réécrit en un chemin qui tient compte de la partie de discours de l'ancre. Notons que les règles 4 et 5 ne sont que des réécritures, leurs probabilités sont donc égales à 1.

La principale limite de cette hypothèse réside dans le fait que les sous-catégorisations n'ayant jamais été observées ne peuvent pas être proposées par l'algorithme. Le corpus d'entraînement étant restreint, il ne serait pas surprenant que d'autres combinaisons puissent exister. L'hypothèse d'indépendance des réalisations syntaxiques de chaque rôle est matérialisée par la règle 8.

3.1.2 G2 : Grammaire ne tenant pas compte des sous-catégorisations

Cette grammaire se distingue de la précédente principalement par le fait qu'elle permet de générer des sous-catégorisations jamais observées dans le corpus d'apprentissage. Il n'y aura donc plus qu'un symbole non terminal par cadre pour représenter sa sous-catégorisation. Les r rôles possibles d'un cadre ne dépendent que de la probabilité d'avoir observé ce rôle avec ce cadre indépendamment des autres rôles qui seraient instanciés. Comme cette grammaire est moins restrictive que la précédente, elle ne comporte "que" 19124 règles de réécriture.

	Probabilité	**Règle**	
1	$P(F_i)$	$\mathcal{F} \rightarrow F_i$	$\forall i \in 1, ..., m$
2	$P(F_{nul})$	$\mathcal{F} \rightarrow F_{nul}$	
3	$P(F_{i,p}\|F_i)$	$F_i \rightarrow F_{i,p}$	$\forall i \in 1, ..., m,\ p \in \{NN, JJ, VB\}$
4	$P(A_j, S_k\|F_{i,p}) = 1$	$F_{i,p} \rightarrow A_j S_i$	$\forall i \in 1, ..., m,\ j \in 1, ..., a_i$
5	$P(A_j\|F_{nul}) = 1$	$F_{nul} \rightarrow A_j$	$\forall j \in 1, ..., a_{nul}$
6	$P(mot\|A_j)$	$A_j \rightarrow mot$	$\forall i \in 1, ..., m,\ j \in 1, ..., a_i$
7	$P(R_x\|S_i)$	$S_i \rightarrow R_x S_i$	$\forall i \in 1, ..., m\ x \in 1, ..., r_i$
8	$P(\varepsilon\|S_i)$	$S_i \rightarrow \varepsilon$	$\forall i \in 1, ..., m$
9	$P(chemin\|R_x, p)$	$R_r \rightarrow chemin$	$\forall x \in 1, ..., r_i,\ i \in 1, ..., m$

Cette fois l'hypothèse est que les rôles sont complètement indépendants les uns des autres. Le nombre de rôles peut donc varier librement.

Par rapport à la grammaire G1, la règle 7 est modifiée, et la 8 est ajoutée. La modification de la règle 7 permet une récursivité dans le choix des rôles, avec la règle 8 mettant un terme à la récursion en réécrivant la sous-catégorisation en ε. Le calcul de la probabilité de la règle 8 a nécessité quelques ajustements, car cette probabilité n'est pas estimable sur le corpus d'entraînement. En effet, elle représente le fait qu'un cadre possède un nombre fini de rôles. Sa probabilité devrait être égale à 1. Sa valeur a été déterminée empiriquement à 0.005. Cette masse de probabilités a été prélevée sur les règles du même type que la règle 7.

La principale limite de cette hypothèse réside dans le fait qu'un même rôle peut être présent plusieurs fois dans une même occurrence de cadre. De même, deux rôles s'excluant l'un l'autre pourraient ainsi cohabiter. Cependant, comme nous l'avons déjà énoncé, cette grammaire a l'avantage de pouvoir créer des sous-catégorisations nouvelles, bien que limitées par leur taille maximale.

3.1.3 G3 : Grammaire tenant compte de la taille des sous-catégorisations

Pour cette grammaire nous allons aussi faire l'hypothèse de l'indépendance des rôles, mais nous allons aussi donner de l'importance à la taille de la sous-catégorisation proposée. Pour cela, on introduit un nouveau type de symbole non terminal : $S_i(t = 1)$. Ce symbole représente les sous-catégorisations de taille 1 pour le cadre i. De même T_i sera la taille maximale des sous-catégorisations d'un cadre i. Cette grammaire, un peu plus coercitive que la précédente mais toujours bien moins que la première, compte 20366 règles de réécriture.

	Probabilité	**Règle**	
1	$P(F_i)$	$\mathcal{F} \rightarrow F_i$	$\forall i \in 1, ..., m$
2	$P(F_{nul})$	$\mathcal{F} \rightarrow F_{nul}$	
3	$P(F_{i,p}\|F_i)$	$F_i \rightarrow F_{i,p}$	$\forall i \in 1, ..., m,\ p \in \{NN, JJ, VB\}$
4	$P(A_j, S_i(t >= 0)\|F_{i,p}) = 1$	$F_{i,p} \rightarrow A_j S_i(t >= 0)$	$\forall i \in 1, ..., m,\ j \in 1, ..., a_i$
5	$P(A_j\|F_{nul}) = 1$	$F_{nul} \rightarrow A_j$	$\forall j \in 1, ..., a_{nul}$
6	$P(mot\|A_j)$	$A_j \rightarrow mot$	$\forall i \in 1, ..., m,\ j \in 1, ..., a_i$
7	$P(R_x, S_i(t >= 1)\|S_i(t >= 0))$	$S_i(t >= 0) \rightarrow R_x S_i(t >= 1)$	$\forall i \in 1, ..., m\ x \in 1, ..., r_i$
8	$P(\varepsilon\|S_i(t >= 0))$	$S_i(t >= 0) \rightarrow \varepsilon$	$\forall i \in 1, ..., m;$
9	$P(R_x, S_i(t >= 2)\|S_i(t >= 1))$	$S_i(t >= 1) \rightarrow R_x S_i(t >= 2)$	$\forall i \in 1, ..., m,\ x \in 1, ..., r_i$
10	$P(\varepsilon\|S_i(t >= 1))$	$S_i(t >= 1) \rightarrow \varepsilon$	$\forall i \in 1, ..., m;$
	$\vdots$		
11	$P(R_x\|S_i(t >= T_i - 1))$	$S_i(t >= T_i - 1) \rightarrow R_x$	$\forall i \in 1, ..., m\ x \in 1, ..., r_i$
12	$P(\varepsilon\|S_i(t >= T_i - 1))$	$S_i(t >= T_i - 1) \rightarrow \varepsilon$	$\forall i \in 1, ..., m;$
13	$P(chemin\|R_x, p)$	$R_x \rightarrow chemin$	$\forall x \in 1, ..., r_i,\ i \in 1, ..., m$

Par rapport à la grammaire précédente, cette grammaire introduit quantité de nouveaux symboles, correspondant aux tailles des sous-catégories en cours de création. Les règles 7 et 8 de G2 deviennent ici une série de règles tenant compte de la taille de la sous-catégorisation actuelle. Nous perdons donc le caractère récursif de la précédente. Les probabilités associées à chaque étape de la grammaire permettent de forcer les sous-catégorisations générées à contenir un nombre de rôles observé dans le corpus d'entraînement.

L'hypothèse principale ici est proche de la précédente : la réalisation d'un rôle est indépendante de celles des autres. On ajoute ici l'hypothèse que la taille des sous-catégorisations est importante. De ce fait, seules les tailles de sous-catégorisations observées dans le corpus d'entraînement seront valables. Cette grammaire étant basée sur la même hypothèse que la précédente, elle souffre des même biais. Ici aussi un même rôle peut être répété plusieurs fois pour une seule occurrence de cadre, et des rôles antagonistes peuvent apparaître simultanément.

3.2 Résultats

Avant de donner les résultats obtenus pour nos différentes grammaires, il nous faut définir quelles sont les mesures que nous allons utiliser pour les comparer et quelle référence nous allons utiliser.

Commençons par la référence. Nous avons créé une référence triviale qui à chaque ancre potentielle repérée assigne le cadre le plus observé pour cette ancre, cadre nul compris. Dans le cas où l'ancre n'a jamais été observée dans le corpus d'entraînement, on attribue à l'ancre le cadre compatible (selon le lexique) le plus observé avec des ancres de cette catégorie. Si jamais à ce stade toujours aucune solution n'a été trouvée, on attribue à l'ancre le cadre nul. Cette référence, contrairement à notre tâche, ne se charge que d'attribuer un cadre à une partie de discours, et ne se préoccupe pas des rôles de ce cadre.

La grammaire va nous permettre de calculer deux probabilités : la probabilité qu'un mot soit associé à un cadre et la probabilité d'une instanciation complète (cadre et sous-catégorisation). La première probabilité est comparable à la référence alors que la seconde ne l'est pas.

Deux types de prédiction sémantique sont faites :
- $\widehat{F} = \underset{F_{SC}}{argmax} P(F_{SC}|S)$ pour la sélection de cadres avec une sous-catégorisation ;
- $\tilde{F} = \underset{F}{argmax} \sum_{SC} P(F_{SC}|S)$ pour la sélection de cadres toutes sous-catégorisations confondues.

A l'issue des traitements on prédit pour un mot :

1. Son cadre le plus probable indépendamment des rôles instanciés ($\tilde{F}$) ;
2. L'instanciation la plus probable (sélection du cadre et des rôles $\widehat{F}$).

Pour mesurer les performances en sélection de cadres indépendamment, nous allons utiliser un *taux de réussite* (exprimé en pourcent). On compte le nombre d'ancres dans le corpus analysé, et le nombre d'ancres dont le cadre a été correctement assigné (cadre nul compris).

Les performances en étiquetage des rôles seront mesurées uniquement parmi les cadres correctement attribués. Nous utiliserons ici trois scores pour mesurer la performance : précision, rappel et F-mesure (tous les trois exprimés en pourcent). La précision mesure, parmi tous les mots qui ont été étiquetés rôles ceux qui l'ont été correctement. Le rappel mesure, parmi tous les mots qui sont des rôles dans la référence combien ont été correctement éiquetés. La F-mesure est une moyenne harmonique des deux scores précédents.

Pour mieux comprendre les forces et les faiblesses de notre automatisation nous allons donner chacune de ces quatre mesures selon la classe syntaxique de l'ancre.

La table 5 permet de comparer les résultats des trois grammaires et de la référence sans tenir compte des rôles (correspond au calcul de $\tilde{F}$).

Partie de discours	Nombre d'occurrences	Référence	G1	G2	G3
Verbes	775	58,32	51,48	53,16	53,55
Adjectifs	660	55,15	48,48	49,09	49,09
Noms	1687	62,89	55,6	58,92	58,86
global	3122	60,12	53,07	55,41	55,48

TABLE 5 – Taux de réussite exprimés en pourcentages, pour la référence et les trois grammaires dans la tâche de sélection de cadres sans tenir compte des rôles

On observe dans la table 5 que la référence est supérieure en tous points à nos grammaires sur la tâche d'assignation de cadres sémantiques (sans sélection des rôles). Les grammaires G2 et G3 donnent les meilleurs résultats. La grammaire G1 ne permet pas de prédire les sous-catégorisations qui n'ont jamais été observées dans le corpus d'entraînement (622 occurrences parmi 4427 dans le corpus de test), ce qui explique probablement ses moindres résultats. La grammaire G2, malgré sa rusticité permet de dépasser les performances de G1, certainement car elle ne se restreint pas aux sous-catégorisations observées dans le corpus d'entraînement. G3 obtient les meilleurs résultats, car à l'instar de G2, elle permet de prédire des sous-catégorisations jamais observées. En revanche, elle introduit un peu plus de contraintes en disqualifiant les sous-catégorisations ayant un nombre de rôles improbable. Cependant la différence entre G2 et G3 n'est pas significative, ce qui laisse penser que l'ajustement de la règle 8 de G2 est suffisante pour contraindre le nombre de sous-catégorisations.
Il est intéressant de noter que les performances dans cette tâche ne sont pas directement liés à l'ambiguïté moyenne de ces parties de discours (cf. table 1).

3.2.1 G1 : Grammaire proche de la sous-catégorisation

La table 5(a) représente les résultats de la grammaire G1 pour la tâche d'analyse sémantique complète (cadre et rôles, équivalent à $\widehat{F}$). On peut remarquer que la précision sur les verbes est assez bonne, les prédictions de rôles faites sur cette catégorie sont donc raisonnables (63,82), dès lors que le cadre a été bien sélectionné. Par contre les résultats sur les adjectifs et noms ont une précision globalement plus faible, probablement du fait que les moyens de réalisation syntaxique des actants sémantiques sont plus limités, d'où plus d'ambiguïté.

3.2.2 G2 : Grammaire ne tenant pas compte des sous-catégorisations

La table 5(b) présente les résultats de la grammaire G2 pour la tâche de sélection des cadres avec rôles. Les performances en sélection de cadres sont légèrement supérieures à celles de la grammaire G1, surtout en ce qui concerne les noms. La précision en sélection de rôles est cependant meilleure pour les verbes et pour les noms. On remarque également que les scores de rappel sont inférieurs à la grammaire G1 : contrairement à ce que nous pensions, cette grammaire sélectionne moins de rôles (rappel bas, notamment pour les verbes : 26,30) que la précédente. Les scores de F-mesure montrent cependant que la grammaire 1 est plus performante pour sélection de rôles.

(a) G1

PoS	Taux de réussite sélection cadre	rôles concernés	Precision rôles	Rappel rôles	F-mesure rôles
Verbes	51,23	723	63,82	43,87	52,00
Adjectifs	47,27	556	59,33	71,58	64,88
Noms	54,71	1768	49,08	57,54	52,97
global	52,27	3047	53,96	56,24	55,08

(b) G2

PoS	Taux de réussite sélection cadre	rôles concernés	Precision rôles	Rappel rôles	F-mesure rôles
Verbes	52,26	736	67,72	26,30	37,89
Adjectifs	47,34	542	59,17	59,64	59,40
Noms	58,49	1690	56,04	51,16	53,49
global	54,59	2968	58,44	45,37	51,08

(c) G3

PoS	Taux de réussite sélection cadre	rôles concernés	Precision rôles	Rappel rôles	F-mesure rôles
Verbes	52,26	735	67,73	25,99	37,57
Adjectifs	47,58	546	58,82	59,74	59,28
Noms	58,51	1689	56,53	51,03	53,64
global	54,64	2970	58,66	45,25	51,09

TABLE 6 – Résultats des grammaires présentées en 3.1 pour la sélection de cadres et de rôles ($\widehat{F}$), en fonction de la partie de discours de l'ancre
La colonne *Taux de réussite sélection cadre* renseigne sur la quantité de cadres correctement attribués, alors que les suivantes sont des statistiques sur les rôles des cadres correctement attribués.

3.2.3 G3 : Grammaire tenant compte de la taille des sous-catégorisations

La table 5(c) représente les résultats pour la grammaire G3 en sélection de cadres complets (cadre et rôles). Bien que meilleure que la grammaire G2, les résultats de ces deux grammaires sont quasiment identiques. Peut-être que la contrainte sur la taille des sous-catégorisations n'est pas suffisante pour creuser l'écart.

4 Conclusion

Les expériences que nous venons de présenter nous montrent que la syntaxe d'une phrase ne permet d'induire que partiellement sa structure sémantique. Nous avons cependant pu remarquer des résultats significativement meilleurs lorsque les ancres des cadres sont des verbes, malgré une ambiguïté sémantique en moyenne plus grande. Ceci s'explique par la richesse des réalisations syntaxiques associées à cette partie de discours : une réalisation syntaxique est toujours associée à un même rôle. Quant aux noms et aux adjectifs, les réalisations syntaxiques qui leurs sont associées sont peu variées : une même réalisation syntaxique peut être associée à plusieurs rôles.
Les grammaires que nous avons présentées ici ne sont pas encore au niveau de notre référence, et bien des améliorations peuvent encore y être apportées. La modification la plus importante à ajouter serait de prendre en compte la nature des mots jouant un rôle dans les cadres. Par exemple pour l'action *manger*, seuls les *êtres vivants* sont capables d'endosser le rôle de *mangeur*. Pour réaliser cette modification, nous pourrions soit faire appel à la sémantique distributionnelle à partir de grands corpus, soit utiliser les ressources réalisées dans des projets comme *WordNet*.
Concernant les améliorations au niveau de la syntaxe, comme nous traitons les cadres indépendamment les uns des autres, nous pourrions utiliser le corpus des phrases d'exemple, annoté partiellement pour la sémantique. Cet ajout permettrait d'enrichir notre corpus d'entraînement et ainsi d'affiner les probabilités de nos grammaires.
Au niveau des données FrameNet, nous pourrions tirer un meilleur parti des informations représentées dans les données. En effet, il existe des relations de plusieurs types entre cadres sémantiques, comme des relations d'héritage ou d'utilisation. En ce qui concerne la première, elle permettrait de pallier le manque de données dans notre corpus de phrases :

les structures syntaxiques d'un cadre père sont proches de celles de ses fils, de plus il existe une table de correspondance entre leurs rôles. Par exemple le cadre *Scrutiny* (prêter attention) est un cadre père de *Research* (rechercher). Grâce à cette méthode, il est possible de regrouper les données des cadres fils au sein d'un même père afin d'augmenter la précision, du moins pour les rôles qui correspondent entre père et fils. La relation d'utilisation est une relation observable dans le corpus, elle consiste à savoir quels cadres sont observés en *symbiose*, c'est à dire qu'un acteur de l'un est un acteur ou une ancre de l'autre. Cette relation, en plus d'être observable, est définie dans les données FrameNet. Elle nous permettrait de créer un modèle identifiant un cadre en fonction des autres cadres de la phrase courante.

Remerciements

Ces travaux sont financés par le projet ANR Asfalda ANR-12-CORD-0023. Nous tenons à remercier Emmanuel Prunet qui nous a fourni une implémentation efficace de l'analyseur d'Earley, ainsi que Marie Candito, Alexis Nasr et Benoit Favre pour leurs relectures et leur encadrement.

Références

BOHNET B. (2010). Very high accuracy and fast dependency parsing is not a contradiction. In *Proceedings of the 23rd International Conference on Computational Linguistics*, p. 89–97 : Association for Computational Linguistics.

BOHNET B., NIVRE J., BOGUSLAVSKY I., GINTER R. F. F. & HAJIC J. (2013). Joint morphological and syntactic analysis for richly inflected languages. *Transactions of the Associtation for Computational Linguistics*, **1**.

COLLINS M. (1997). Three generative, lexicalised models for statistical parsing. In *Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and Eighth Conference of the European Chapter of the Association for Computational Linguistics*, p. 16–23 : Association for Computational Linguistics.

DAS D., SCHNEIDER N., CHEN D. & SMITH N. A. (2010). Probabilistic frame-semantic parsing. In *Human language technologies : The 2010 annual conference of the North American chapter of the association for computational linguistics*, p. 948–956 : Association for Computational Linguistics.

DAS D. & SMITH N. A. (2011). Semi-supervised frame-semantic parsing for unknown predicates : Supplementary material. In *Proc. of ACL*.

EARLEY J. (1970). An efficient context-free parsing algorithm. *Communications of the ACM*, **13**(2), 94–102.

FILLMORE C. (1982). Frame semantics. *Linguistics in the morning calm*, p. 111–137.

FILLMORE C. J. (1976). Frame semantics and the nature of language*. *Annals of the New York Academy of Sciences*, **280**(1), 20–32.

FILLMORE C. J. (1977). The case for case reopened. *Syntax and semantics*, **8**(1977), 59–82.

FILLMORE C. J. (1985). Frames and the semantics of understanding. *Quaderni di semantica*, **6**(2), 222–254.

FILLMORE C. J. & BAKER C. (2010). A frames approach to semantic analysis. *The Oxford handbook of linguistic analysis*, p. 313–339.

FILLMORE C. J. & BAKER C. F. (2001). Frame semantics for text understanding. In *Proceedings of WordNet and Other Lexical Resources Workshop, NAACL*.

MILLER G. & FELLBAUM C. (1998). Wordnet : An electronic lexical database.

MODI A., TITOV I. & KLEMENTIEV A. (2012). Unsupervised induction of frame-semantic representations. In *Proceedings of the NAACL-HLT Workshop on the Induction of Linguistic Structure*, p. 1–7 : Association for Computational Linguistics.

PALMER M., GILDEA D. & KINGSBURY P. (2005). The proposition bank : An annotated corpus of semantic roles. *Computational Linguistics*, **31**(1), 71–106.