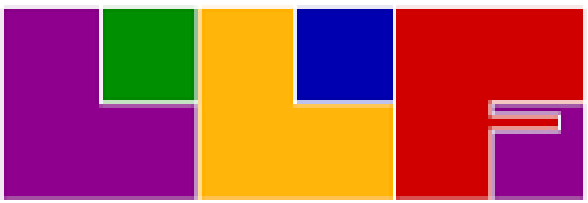
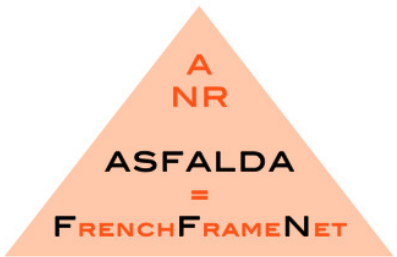
Représentations sémantiques à la FrameNet : création d'un FrameNet pour le français et utilisation en apprentissage supervisé

Marie Candito,
en collaboration avec
Marianne Djemaa, Philippe Muller, Laure Vieu,
G. de Chalendar, B. Sagot, P. Amsili (pour FrameNet FR)
C. Ribeyre, D. Seddah, G. Perrier, B. Guillaume, K. Fort (syntaxe profonde)
Olivier Michalon, Alexis Nasr (analyse sémantique)

Séminaire LIF
26 avril 2017





Plan

1. Construction d'un FrameNet du français
   - qu'est-ce que FrameNet et pourquoi un FrameNet FR ?
   - méthodologie
   - évaluation
   - état de la ressource
2. Annotation de type FrameNet : retours d'expérience
3. Parsing FrameNet: impact d'une syntaxe normalisée

1. Construction d'un FrameNet du français

Projet ANR ASFALDA : Objectif

Développer des ressources et outils pour l'analyse sémantique de surface du français

- Cadre choisi = FrameNet (Baker et al. 1998; Fillmore 2007)
  - projet d'annotation en rôle sémantiques en anglais
  - instanciant la "frame semantics"
  - développé à l'ICSI à Berkeley (depuis 1998...)

FrameNet : caractéristiques

"Instantiation" pratique et à grande échelle de la frame semantics de C. Fillmore

- frame = situation prototypique
  - évocable via des unités lexicales ($\longrightarrow$ les déclencheurs)
    - par ex: frame Commitment, évocable par *promettre*, *promesse*, *s'engager* ...
  - et précisant les participants et propriétés intervenant dans le frame ($\longrightarrow$ les remplisseurs de rôles)

Le ministre français ***S'*** *est* ***ENGAGÉ*** *auprès_du président de l' Ukraine (...) à transférer gratuitement les codes de calcul*

Commitment; Speaker, Addressee, Message

FrameNet : caractéristiques

Structure: frames et relations entre frames

- un peu plus de 1000 frames
- reliés par relations (héritage, perspective...)

Lexique anglais

- environ 12000 “sens” : couples lemme/frame

Annotations

- lexicographiques : exemples à annoter (BNC), choisis
  - pour représenter la variété des réalisations syntaxiques des rôles
  - pour dériver généralisations de “linking”
- full-text : annotation complète de textes
  - plus adapté pour de l’apprentissage automatique
  - (cf. préservation de la distribution des sens / des réalisations syntaxiques)

FrameNet : caractéristiques

Structure: frames et relations entre frames

- un peu plus de 1000 frames
- reliés par relations (héritage, perspective...)

Lexique anglais

- environ 12000 "sens" : couples lemme/frame

Annotations

- lexicographiques : exemples à annoter (BNC), choisis
  - pour représenter la variété des réalisations syntaxiques des rôles
  - pour dériver généralisations de "linking"
- full-text : annotation complète de textes
  - plus adapté pour de l'apprentissage automatique
  - (cf. préservation de la distribution des sens / des réalisations syntaxiques)

FrameNet : caractéristiques

Structure: frames et relations entre frames

- un peu plus de 1000 frames
- reliés par relations (héritage, perspective...)

Lexique anglais

- environ 12000 "sens" : couples lemme/frame

Annotations

- lexicographiques : exemples à annoter (BNC), choisis
  - pour représenter la variété des réalisations syntaxiques des rôles
  - pour dériver généralisations de "linking"
- full-text : annotation complète de textes
  - plus adapté pour de l'apprentissage automatique
  - (cf. préservation de la distribution des sens / des réalisations syntaxiques)

FrameNet : caractéristiques
Des rôles de granularité variable

- Les rôles dans FrameNet sont spécifiques à un frame
- (pas de liste réduite de rôles sémantiques)
- Mais rôles généralisés dérivables via relations entre frames
- $\longrightarrow$ indirectement: on peut choisir un niveau de granularité des rôles

- FR_Statement hérite de Communication
  - le rôle Speaker de FR_Statement correspond au rôle Communicator de Communication

FrameNet : caractéristiques
Des frames de granularité variable

- un frame regroupe plusieurs lexèmes
  - éventuellement de catégories diverses
    - frame Causation: déclencheurs *parce que*, *tant et si bien que*
    - *causer*, *résulter*, ...
    - *cause*, *conséquence*, *retombée* ...
    - *à cause de*, *grâce à*, ...

- volonté d'avoir des critères prioritairement sémantiques (≠classes de Levin et Verbnet, classes LADL...)
  - ⇒ caractère partiellement arbitraire de la granularité sémantique des frames
  - ⇒ on peut moins s'appuyer sur les critères formels

Motivations

Pas de ressource Fr générale annotée en rôles sémantiques

Choix de réutiliser la ressource existante FrameNet:

- cf. généralisations lexicales : les frames groupent plusieurs lexèmes ($\neq$ PropBank)
- cf. critères non prioritairement syntaxiques ($\neq$ VerbNet)
  - portabilité Boas, 2009
  - généralisation : les frames peuvent grouper différentes catégories de lexèmes
- Choix d'annoter en corpus, en respectant la distribution naturelle des phénomènes linguistiques

Le projet ASFALDA en bref

Projet ANR: oct 2012 à juin 2016
Partenaires: Alpage, CEA-List, LIF, LLF, MELODI (IRIT), Ant'inno

1. Annotation d'un corpus en frames et en rôles
   - 1.1 définition de la couverture visée: en termes de domaines notionnels
   - 1.2 Modélisation en frames de ces domaines
     - en partant des frames de l'anglais
   - 1.3 Construction du lexique français associé
   - 1.4 Annotation sur corpus arborés
     - FrenchTreebank + SequoiaTreebank
   - 1.5 Extraction du lexique d'après annotations
     - informations de fréquence
     - patrons de linking

Le projet ASFALDA en bref

Projet ANR: oct 2012 à juin 2016
Partenaires: Alpage, CEA-List, LIF, LLF, MELODI (IRIT), Ant'inno

1. Annotation d'un corpus en frames et en rôles
   - 1.1 définition de la couverture visée: en termes de domaines notionnels
   - 1.2 Modélisation en frames de ces domaines
     - en partant des frames de l'anglais
   - 1.3 Construction du lexique français associé
   - 1.4 Annotation sur corpus arborés
     - FrenchTreebank + SequoiaTreebank
   - 1.5 Extraction du lexique d'après annotations
     - informations de fréquence
     - patrons de linking

Le projet ASFALDA en bref

Projet ANR: oct 2012 à juin 2016
Partenaires: Alpage, CEA-List, LIF, LLF, MELODI (IRIT), Ant'inno

1. Annotation d'un corpus en frames et en rôles
   - 1.1 définition de la couverture visée: en termes de domaines notionnels
   - 1.2 Modélisation en frames de ces domaines
     - en partant des frames de l'anglais
   - 1.3 Construction du lexique français associé
   - 1.4 Annotation sur corpus arborés
     - FrenchTreebank + SequoiaTreebank
   - 1.5 Extraction du lexique d'après annotations
     - informations de fréquence
     - patrons de linking

Le projet ASFALDA en bref

Projet ANR: oct 2012 à juin 2016
Partenaires: Alpage, CEA-List, LIF, LLF, MELODI (IRIT), Ant'inno

1. Annotation d'un corpus en frames et en rôles
   - 1.1 définition de la couverture visée: en termes de domaines notionnels
   - 1.2 Modélisation en frames de ces domaines
     - en partant des frames de l'anglais
   - 1.3 Construction du lexique français associé
   - 1.4 Annotation sur corpus arborés
     - FrenchTreebank + SequoiaTreebank
   - 1.5 Extraction du lexique d'après annotations
     - informations de fréquence
     - patrons de linking

Le projet ASFALDA en bref

Projet ANR: oct 2012 à juin 2016
Partenaires: Alpage, CEA-List, LIF, LLF, MELODI (IRIT), Ant'inno

1. Annotation d'un corpus en frames et en rôles
   - 1.1 définition de la couverture visée: en termes de domaines notionnels
   - 1.2 Modélisation en frames de ces domaines
     - en partant des frames de l'anglais
   - 1.3 Construction du lexique français associé
   - 1.4 Annotation sur corpus arborés
     - FrenchTreebank + SequoiaTreebank
   - 1.5 Extraction du lexique d'après annotations
     - informations de fréquence
     - patrons de linking

Le projet ASFALDA en bref

Projet ANR: oct 2012 à juin 2016
Partenaires: Alpage, CEA-List, LIF, LLF, MELODI (IRIT), Ant'inno

1. Annotation d'un corpus en frames et en rôles
   - 1.1 définition de la couverture visée: en termes de domaines notionnels
   - 1.2 Modélisation en frames de ces domaines
     - en partant des frames de l'anglais
   - 1.3 Construction du lexique français associé
   - 1.4 Annotation sur corpus arborés
     - FrenchTreebank + SequoiaTreebank
   - 1.5 Extraction du lexique d'après annotations
     - informations de fréquence
     - patrons de linking

Le projet ASFALDA en bref

2. Automatisation
   - 2.1 Construction d'un analyseur sémantique associé
     - (LIF, CEA-List, Alpage)
   - 2.2 Application
     - Intégration au moteur de recherche d'information CEA-List
     - et au logiciel Ant'Box de la société Ant'inno

2. Structure de frames & lexique

Structure de frames

Focalisation sur 4 domaines notionnels :

- 3 domaines spécifiques
  - transactions commerciales
  - communication verbale
  - positions cognitives
- 1 domaine transversal:
  - causalité

Structure de frames & lexique

Travail itératif sur frames et lexique (par les 10 experts):

- Sélection/adaptation au français des frames anglais
- Construction du lexique Fr associé, en utilisant:
  - lexiques FN-Fr automatiquement obtenus à partir du lexique FN-En (Padò, 2007; Mouton et al., 2010)
  - lexiques Fr généraux ou spécifiques (tables LADL, LVF, LexConn (Roze et al., 12), Casoar (Benamara et al., 11), French-TimeBank (Bittar, 10), le Dictionnaire Electronique des Synonymes)
  - concordancier sur les corpus visés

Structure de frames & lexique: retours d'expérience

Différentes stratégies de travail:

- Berkeley FrameNet: frame par frame
- SALSA (projet allemand, Burchardt et al. 2006) : lemme par lemme
- ASFALDA: domaine par domaine
  - couverture lexicale exhaustive pour certains domaines
  - limitation de la polysémie

Retours d'expérience:

- Délimitation du périmètre sémantique des frames: très difficile
  - difficultés à comprendre les distinctions de frames
  - Introduction de critères syntaxiques plus précis pour décider quand distinguer 2 frames
- Fusions de frames / adaptation de frames (41 frames sur 105)
  - A comparer aux 33% pour SALSA (projet FrameNet allemand)
- Plus 20 nouveaux frames (pour couverture des domaines)

Structure de frames & lexique: retours d'expérience

Différentes stratégies de travail:

- Berkeley FrameNet: frame par frame
- SALSA (projet allemand, Burchardt et al. 2006) : lemme par lemme
- ASFALDA: domaine par domaine
  - couverture lexicale exhaustive pour certains domaines
  - limitation de la polysémie

Retours d'expérience:

- Délimitation du périmètre sémantique des frames: très difficile
  - difficultés à comprendre les distinctions de frames
  - Introduction de critères syntaxiques plus précis pour décider quand distinguer 2 frames
- Fusions de frames / adaptation de frames (41 frames sur 105)
  - A comparer aux 33% pour SALSA (projet FrameNet allemand)
- Plus 20 nouveaux frames (pour couverture des domaines)

3. Annotation en corpus

Annotation en corpus

- 6 annotateurs
- plus “experts” M. Djemaa, P. Muller, L. Vieu et M. Candito

Restriction à 4 domaines:

- Transactions commerciales
- Positions cognitives
- Causalité
- Communication verbale

Annotation en corpus

- Annotation sur des corpus déjà annotés en syntaxe:
  - pour accélérer l'annotation
  - pour extraction de patrons syntaxico-sémantiques

- Corpus utilisés
  - French Treebank (Abeillé and Barrier, 2004)
    - 18535 phrases (version SPMRL 2013 (Seddah et al. 2013))
  - Sequoia Treebank (Candito and Seddah, 2012)
    - 3099 phrases
    - médical, Europarl, Est Républicain, FrWiki
  - environ 21,500 phrases, 625,000 tokens
  - convertis en arbres de dépendances

Annotation en corpus

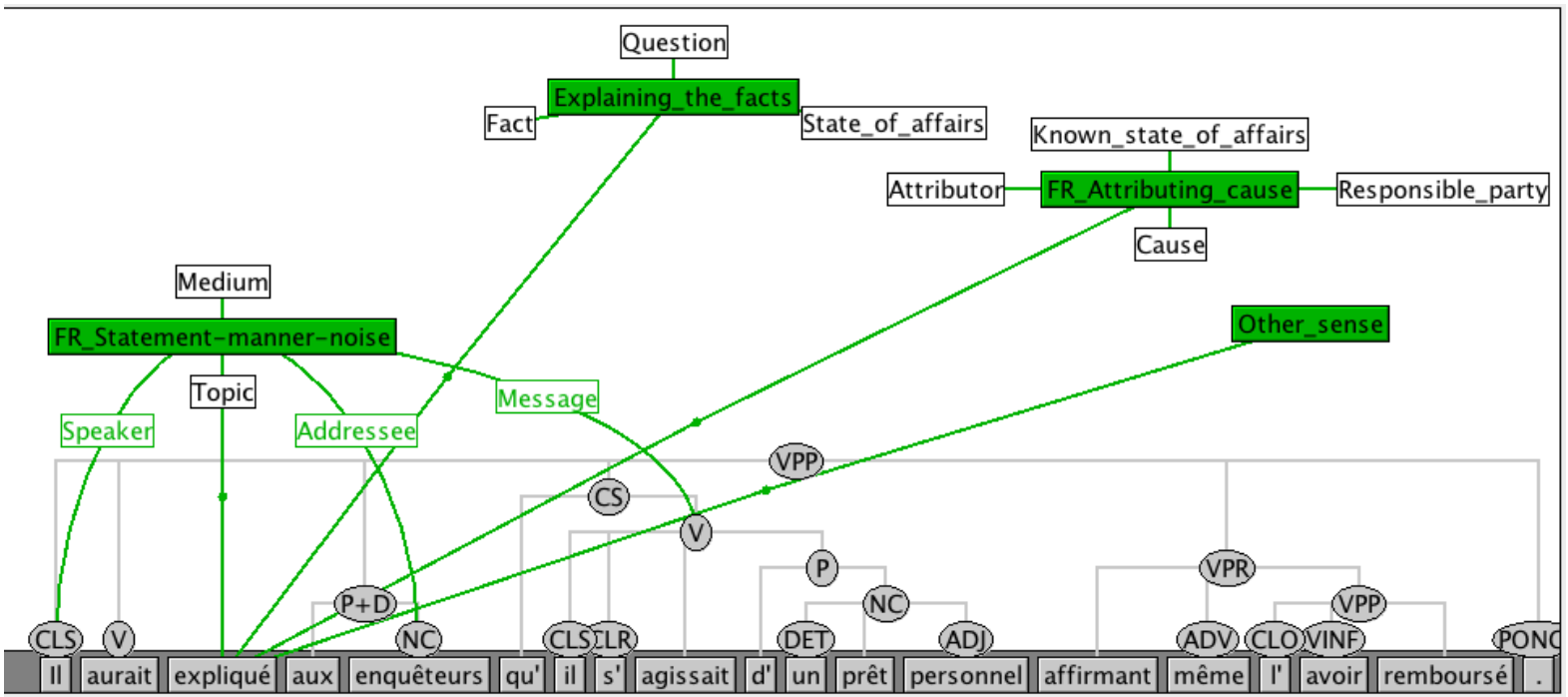
- Annotation sur des corpus déjà annotés en syntaxe:
  - pour accélérer l'annotation
  - pour extraction de patrons syntaxico-sémantiques

- Corpus utilisés
  - French Treebank (Abeillé and Barrier, 2004)
    - 18535 phrases (version SPMRL 2013 (Seddah et al. 2013))
  - Sequoia Treebank (Candito and Seddah, 2012)
    - 3099 phrases
    - médical, Europarl, Est Républicain, FrWiki

  - environ 21,500 phrases, 625,000 tokens
  - convertis en arbres de dépendances

Annotation via outil graphique

- annotation lemme par lemme
- au maximum 100 premières occurrences d'un lemme
- pré-annotation automatique des frames du lemme
  - outil Salto préexistant (Burchardt et al., 2006)
- 2 annotations indépendantes + adjudication



Evaluation: accord inter-annotateur

Entre 2 annotations indépendantes (75% des annotations):

- pour une occ. de déclencheur: Fscore du choix de frame
- pour un frame commun choisi: Fscore du choix des remplisseurs des rôles sémantiques

	Nb d'occ. de déclencheurs	% de Noms	% de Verbes	Fscore inter-annotateur		
				Frame	Exact Role	Partial Role
	17667	36	50	85.9	77.2	81.9
Break-down par domaine						
Commercial	3307	60	40	92.0	73.4	80.4
Causalité	7691	30	48	79.2	74.2	80.4
Pos. Cognitives	7886	28	62	90.6	81.1	86.0
Communication	2221	23	76	89.6	82.3	87.5
Break-down par catégorie du déclencheur						
V	8834	-	-	87.6	82.8	87.1
N	6234	-	-	86.8	68.3	72.5
other	2509	-	-	77.7	74.6	82.1

Livraison 1.2 des annotations sémantiques

http://asfalda.linguist.univ-paris-diderot.fr/frameIndex.xml

- $\longrightarrow$ env. 16200 instances de frames annotées
- $+$ 8750 occ. annotées comme hors domaines (Other_sense)

	Nb frames distincts	Nb déclencheurs distincts	Nb sens	Nb frames annotés ($\neq$ Other_sense)
ALL	105	873	1109	16167
Commercial	19	90	99	2930
Causalité	11	243	285	3895
Positions cog.	44	372	442	5426
Communication	47	347	411	5233
N	-	296	346	5282
V	-	446	594	9165
PREP	-	35	43	674
ADV	-	26	42	407
CONJ	-	22	28	301
ADJ	-	43	48	234

Rôles

- 421 rôles définis au niveau des frames
- regroupés en 40 “macro-roles” définis au niveau du domaine (causalité, comm. lang etc...)

4. Annotation de type FrameNet en corpus: retours d'expérience

Annotation FrameNet en corpus: problèmes typiques

FrameNet original: exemples choisis $\longrightarrow$ Pb d'annotation en corpus assez peu documentés.

L'annotation en corpus recquiert de choisir des solutions précises pour des problèmes typiques:

- d'interface morpho-syntaxe / sémantique
  - statut du clitique "se"
  - annotation des constructions à attribut de l'objet (Djemaa, 2014)
- de sémantique lexicale:
  - identification de la polysémie
  - nominalisations référant à un participant
- divergences entre expression littérale et interprétation
  - expressions figées, métaphores, ellipses...

Focus: gestion des verbes pronominaux

Question: pour une occ. de "se V", annoter avec même frame que "V" ou pas?

- Classification manuelle préalable des occurrences de "se"
  - définition des classes dans le cadre de l'annotation en syntaxe profonde Deep Sequoia (Candito et al., 2014)
  - tâche d'annotation très difficile $\Rightarrow$ adju à 4 (avec M. Djemaa, V. Mouilleron, B. Sagot)
- 4 Classes principales:
  1. vrai réfléchi ou réciproque (*ils se soignent*)
  2. se-moyen (*ça se casse pas avec un marteau*)
  3. se-neutre (*le vase s'est cassé*)
  4. réfléchi intrinsèque (*elle s'aperçoit que...*)
- très très laborieux...

Focus: gestion des verbes pronominaux

Principe FrameNet: stabilité du nombre d'actants sémantiques au sein d'un frame

$\longrightarrow$ déclenchement de frame par le V seul (sans le se) pour:

- vrai réfléchi/réciproque (approximation: le sujet joue 2 rôles sém.)
  - *Les actionnaires doivent se **CONSIDÉRER** comme responsables*
  - *elle a tendance à nous **CONSIDÉRER** comme des mineurs*

- se-moyen (cf. agent sémantiquement présent "fantôme")
  - *La spéculation , il est vrai , ne s' **ANTICIPE** pas*
  - *Les torréfacteurs **ANTICIPENT** un redressement*

Focus: gestion des verbes pronominaux

Principe FrameNet: stabilité du nombre d'actants sémantiques au sein d'un frame

$\longrightarrow$ déclenchement de frame par le V seul (sans le se) pour:

- vrai réfléchi/réciproque (approximation: le sujet joue 2 rôles sém.)
  - *Les actionnaires doivent se **CONSIDÉRER** comme responsables*
  - *elle a tendance à nous **CONSIDÉRER** comme des mineurs*

- se-moyen (cf. agent sémantiquement présent "fantôme")
  - *La spéculation , il est vrai , ne s' **ANTICIPE** pas*
  - *Les torréfacteurs **ANTICIPENT** un redressement*

Focus: gestion des verbes pronominaux

Principe FrameNet: stabilité du nombre d'actants sémantiques au sein d'un frame

$\longrightarrow$ déclenchement de frame par le se+V pour:

- réfléchis intrinsèques
  - *On **S'** en **APERÇOIT** aux lettres que nous recevons*
- se-neutre
  - cf. un participant sémantiquement "effacé" (agent absent)
  - *Cela **S'APPELLE** la politique de la confiance*

Focus: gestion des verbes pronominaux

Principe FrameNet: stabilité du nombre d'actants sémantiques au sein d'un frame

$\longrightarrow$ déclenchement de frame par le se+V pour:

- réfléchis intrinsèques
  - *On **S'** en **APERÇOIT** aux lettres que nous recevons*
- se-neutre
  - cf. un participant sémantiquement "effacé" (agent absent)
  - *Cela **S'APPELLE** la politique de la confiance*

Focus: nom prédicatif référant à un participant

Noms relationnels:

*Reste à déterminer les **CAUSES** exactes de l'incendie*

SYNTAXIQUEMENT:

- *causes* est monovalent: un complément en "de"

SÉMANTIQUEMENT: prédicat sémantique

- monovalent entité_étant_la_cause_de(effet)
- $\longrightarrow$ bivalent : relation_de_cause_à_effet(cause, effet)

(frame Causation; Cause, Effect)

*Reste à déterminer les **CAUSES** exactes de l'incendie*

Focus: nom prédicatif référant à un participant

Noms relationnels:

*Reste à déterminer les **CAUSES** exactes de l'incendie*

SYNTAXIQUEMENT:

- *causes* est monovalent: un complément en "de"

SÉMANTIQUEMENT: prédicat sémantique

- monovalent entité_étant_la_cause_de(effet)
- $\longrightarrow$ bivalent : relation_de_cause_à_effet(cause, effet)

(frame Causation; Cause, Effect)

*Reste à déterminer les **CAUSES** exactes de l'incendie*

Focus: nom prédicatif référant à un participant

Noms relationnels:

*Reste à déterminer les **CAUSES** exactes de l'incendie*

SYNTAXIQUEMENT:

- *causes* est monovalent: un complément en "de"

SÉMANTIQUEMENT: prédicat sémantique

- monovalent entité_étant_la_cause_de(effet)
- $\longrightarrow$ bivalent : relation_de_cause_à_effet(cause, effet)

(frame Causation; Cause, Effect)

*Reste à déterminer les **CAUSES** exactes de l'incendie*

Focus: nom prédicatif référant à un participant

Noms relationnels:

*Reste à déterminer les **CAUSES** exactes de l'incendie*

SYNTAXIQUEMENT:

- *causes* est monovalent: un complément en "de"

SÉMANTIQUEMENT: prédicat sémantique

- monovalent entité_étant_la_cause_de(effet)
- $\longrightarrow$ bivalent : relation_de_cause_à_effet(cause, effet)

(frame Causation; Cause, Effect)

*Reste à déterminer les **CAUSES** exactes de l'incendie*

Prédicat référant à un participant: traitement ASFALDA

Pour ce type de déclencheur, on distingue

- emplois référentiels (majorité des occurrences)
- emplois prédicatifs du déclencheur
  - $\longrightarrow$ on considère que le déclencheur joue son rôle de prédicat sans référer au participant

Distinction valable pour tous les noms:

- *As-tu vu la licorne?*
- *Cet animal est une licorne.*

Prédicat référant à un participant: traitement ASFALDA des emplois référentiels

(frame Causation; Cause, Effect)

*Reste à déterminer les **CAUSES** exactes de l'incendie*

*Les **CONSÉQUENCES** de cette tension sont assez désastreuses*

Prédicat référant à un participant: traitement ASFALDA des emplois prédicatifs

Cas typiques d'emplois prédicatifs:

(frame Causation; Cause, Effect)

- **apposition**:

  *Première **CAUSE** de chômage , les fins de CDD sont en hausse.*

- **phrase copulative**:

  *Cette mesure est la **CONSÉQUENCE** d'une conjoncture toujours défavorable.*

- **phrase copulative inversée**:

  *Le **RÉSULTAT** de ces dispositions est de ne combler qu' une partie du déficit.*

Ambiguïté evt/participant

Cas typique d'un prédicat pouvant référer à un rôle: nominalisations ambigues évènement/participant (ou résultat).

(Commerce_buy; Buyer, Goods)

- *(...) ses **ACHATS** de pièces_détachées automobiles (...)*
  - évènement, ne réfère pas à un rôle mais à l'acte d'achat

- *Le marché des cadeaux - souvenirs et "**ACHATS** à rapporter" (...) a atteint 10,4 milliards de francs.*
  - *achats* réfère aux biens achetés (rôle Goods)

Bilan

Création d'un Framenet du français

- librement disponible
  - (licence nécessaire pour strates morpho-syntaxiques du FTB)
- annotations en corpus ($\sim$ 16000 frames annotés)
- lexique quantifié extrait des annotations ($\sim$ 870 lemmes)
  - dont 490 complètement couverts (pas de "Other_sense")
- faisable avec un accord inter-annotateur satisfaisant
- très coûteux en temps
  - en particulier: délimitation des frames, gestion polysémie
- limitation à 4 domaines, mais complètement couverts
- portabilité multilingue:
  - 39% de frames modifiés (liens conservés avec FN anglais)
  - 19% de frames nouveaux

Bilan

Création d'un Framenet du français

- librement disponible
  - (licence nécessaire pour strates morpho-syntaxiques du FTB)
- annotations en corpus ($\sim$ 16000 frames annotés)
- lexique quantifié extrait des annotations ($\sim$ 870 lemmes)
  - dont 490 complètement couverts (pas de "Other_sense")
- faisable avec un accord inter-annotateur satisfaisant
- très coûteux en temps
  - en particulier: délimitation des frames, gestion polysémie
- limitation à 4 domaines, mais complètement couverts
- portabilité multilingue:
  - 39% de frames modifiés (liens conservés avec FN anglais)
  - 19% de frames nouveaux

5. Analyse automatique FrameNet

Analyse automatique FrameNet

- aspect WSD : quel frame pour une occ. de prédicat ambigu
- aspect SRL : identification des rôles sémantiques

Défis

- généralisation des données
  - WordNet (e.g. Johansson et Nugues, 2007)
  - représentations distribuées (e.g. Hermann et al. ACL 2014)
- modèles joints (choix du frame, des rôles) (Das et al. 2010, 2011)

Analyse automatique FrameNet: quelle syntaxe?

Travail avec **Olivier Michalon**, Corentin Ribeyre, Alexis Nasr
(Michalon et al. Coling 2016)

- traits syntaxiques connus comme décisifs pour le SRL (since Gildea et Jurafsky, 2002)
- cf. régularités de "linking" : réalisations syntaxiques des arguments sémantiques
- perturbées par variation syntaxique
  - changements de diathèse
  - arguments non réalisés localement au prédicat

- $\longrightarrow$ on dispose justement de représentations avec variation syntaxique neutralisée
  - projet Deep Sequoia
  - Alpage / Sémagramme (G. Perrier, B. Guillaume, K. Fort, D. Seddah)
- $\longrightarrow$ étude de l'impact sur le parsing FrameNet

Analyse automatique FrameNet: quelle syntaxe?

Travail avec **Olivier Michalon**, Corentin Ribeyre, Alexis Nasr
(Michalon et al. Coling 2016)

- traits syntaxiques connus comme décisifs pour le SRL (since Gildea et Jurafsky, 2002)
- cf. régularités de "linking" : réalisations syntaxiques des arguments sémantiques
- perturbées par variation syntaxique
  - changements de diathèse
  - arguments non réalisés localement au prédicat

- $\longrightarrow$ on dispose justement de représentations avec variation syntaxique neutralisée
  - projet Deep Sequoia
  - Alpage / Sémagramme (G. Perrier, B. Guillaume, K. Fort, D. Seddah)
- $\longrightarrow$ étude de l'impact sur le parsing FrameNet

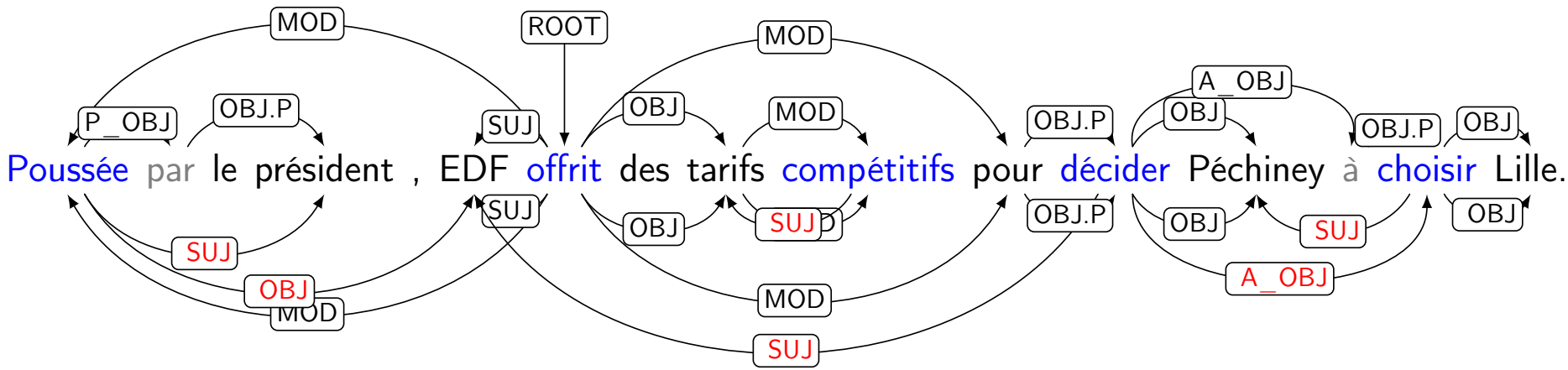
Syntaxe de surface versus syntaxe "profonde"

(Candito et al., 14; Perrier et al. 14)

- arguments non réalisés localement au prédicat
  - "sujet" des verbes non conjugués
    - *Paul veut partir*
    - *les gens aimant la musique / nés en 45 / embauchés en 45*
  - arguments partagés par verbes coordonnés
    - *Paul veut aimé et être aimé*
    - *Paul dort/cuit et vend des crèpes*
- changements de diathèse
- court-circuitage de marqueurs syntaxiques (preps régies...)
- $\longrightarrow$ graphes

Syntaxe de surface versus syntaxe “profonde”

(sans arcs pour déterminants ni ponctuation)

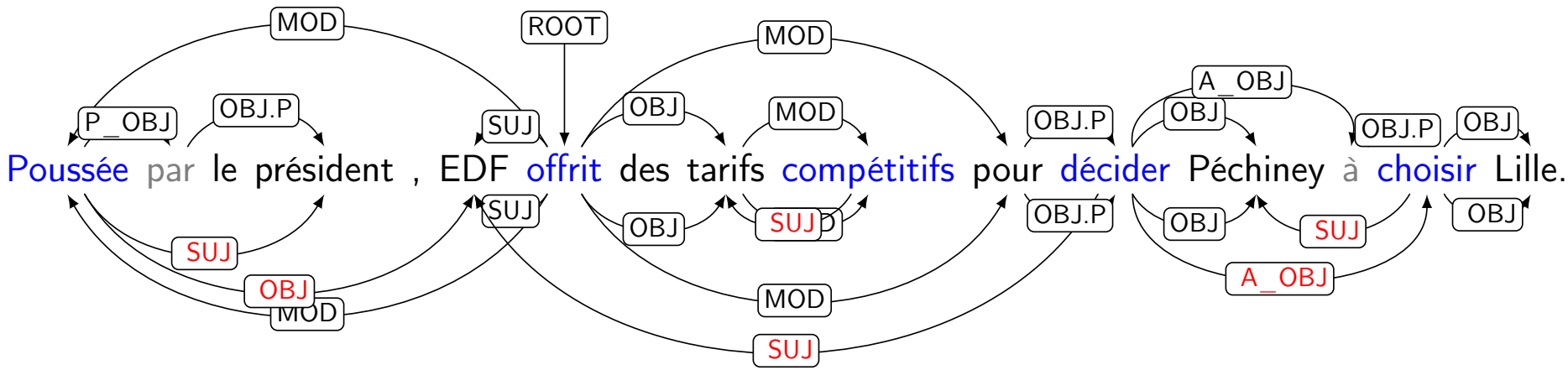


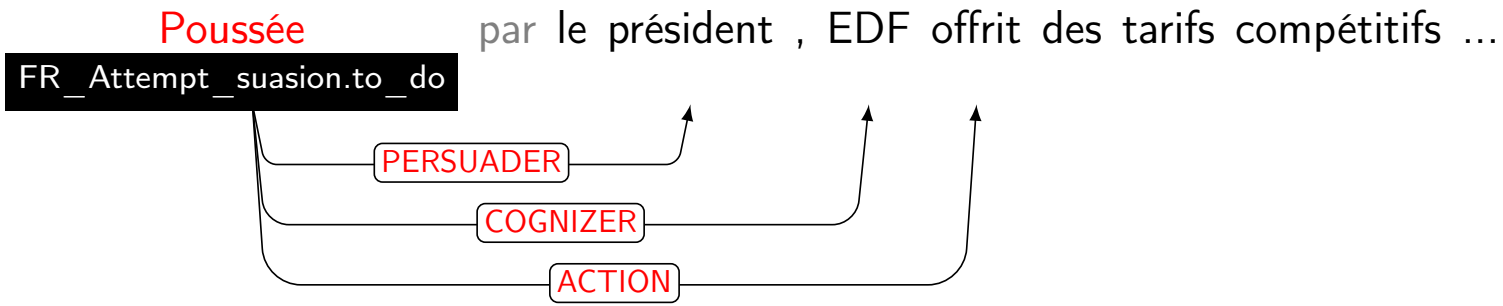
Chemin syntaxique classique entre “Poussée” et “EDF” : -mod,+suj

Chemin syntaxique profond entre “Poussée” et “EDF” : +obj

Syntaxe de surface versus syntaxe “profonde”

(sans arcs pour déterminants ni ponctuation)



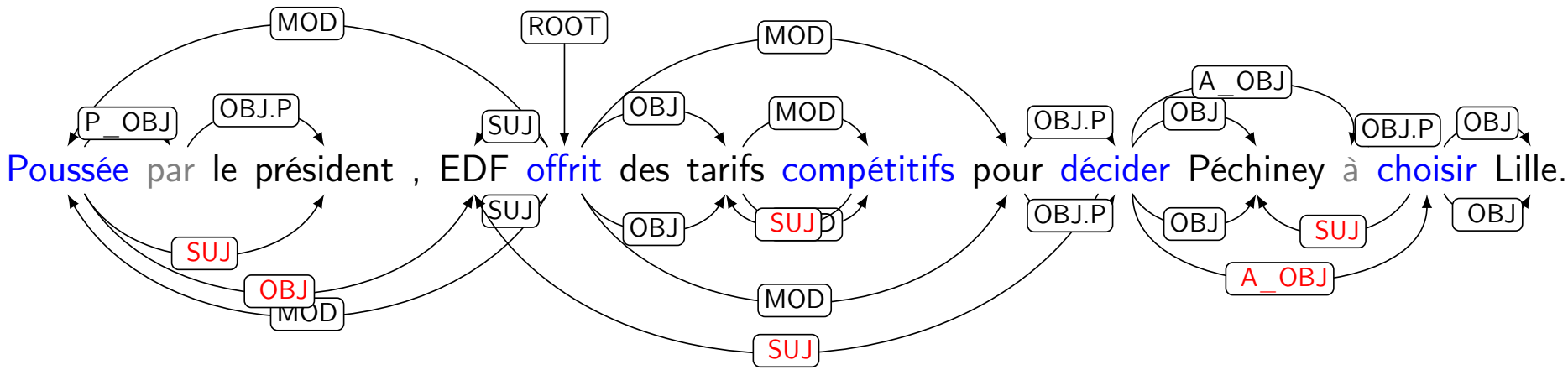


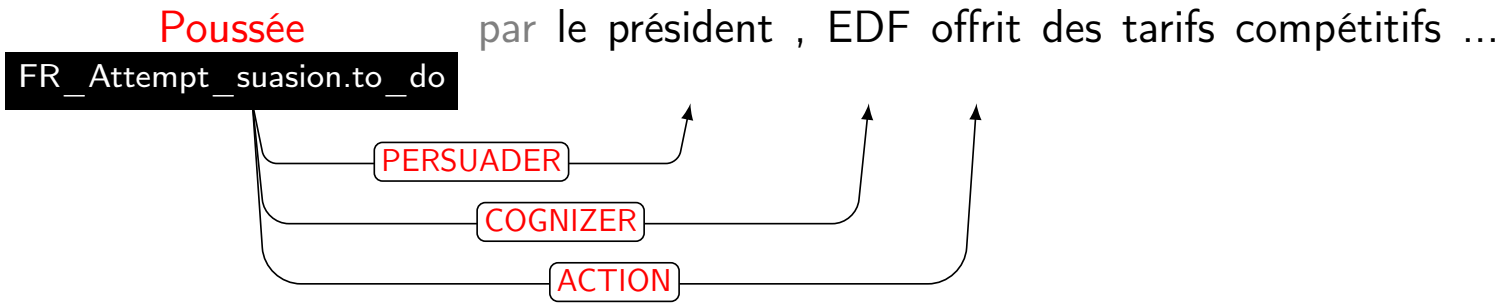
Chemin syntaxique classique entre “Poussée” et “EDF” : -mod,+suj

Chemin syntaxique profond entre “Poussée” et “EDF” : +obj

Syntaxe de surface versus syntaxe “profonde”

(sans arcs pour déterminants ni ponctuation)





Chemin syntaxique classique entre “Poussée” et “EDF” : -mod,+suj

Chemin syntaxique profond entre “Poussée” et “EDF” : +obj

Syntaxe profonde

Données gold: annotation validée manuellement sur corpus Sequoia (3099 phrases)

Données pseudo-gold: règles déterministes de transformation de graphes appliquée au French Treebank (Ribeyre et al, 2014)

Parsing en syntaxe profonde:

- pipeline parsing de surface + règles (baseline)
- voir aussi apprentissage direct d'un parser de graphes (Ribeyre et al., 2015)

Mesure de la normalisation syntaxique

Traits syntaxiques classiques =
chemins syntaxiques entre

- un prédicat
- (la tête syntaxique d') un remplisseur de rôle

Les chemins en syntaxe profonde sont plus réguliers

cf. Entropie moyenne des distributions
P(chemin vers remplisseur de rôle | rôle sémantique) diminue:

- 1.65 avec chemins syntaxiques "classiques"
- 1.32 avec chemins syntaxiques "profonds"

Mesure de la normalisation syntaxique

5 chemins les plus fréquents,
pour les rôles des déclencheurs verbaux

surface syntax		deep syntax	
(+suj)	25.02%	(+suj)	33.10%
(+obj)	17.01%	(+obj)	32.79%
(-mod)	8.04%	(+a_obj)	4.73%
(+obj,+obj.cpl)	4.42%	(-mod)	3.15%
(+a_obj,+obj.p)	4.09%	(+mod,+obj.p)	2.46%
Total	58.58 %	Total	76.23 %

Impact sur le parsing FrameNet

Système basique (pipeline WSD + SRL, classification supervisée)

- WSD : un classifieur par lemme déclencheur
- SRL : un classifieur par frame

Impact positif pour le SRL FrameNet, en particulier pour les déclencheurs verbaux

	Prec.		Recall		F-measure	
Input syntax	surf	deep	surf	deep	surf	deep
WSD (gold frame ≠ Other_sense)	80.1	80.7	80.1	80.7	80.1	80.7
SRL (for gold role filler heads)	81.4	86.4	59.1	66.1	**68.5**	**74.9**

Prec.		Recall		F-measure	
surf	deep	surf	deep	surf	deep
80	80.5	80.8	80.9	80.4	80.7
75.7	80.3	51.6	59.0	**61.3**	**68.0**

Table: FastSem results for **verbs**, using **gold** (top) and **predicted** (bottom) surf and deep syntax.

Perspectives

- Saturation des frames annotés
  - bcp de rôles non instanciés
  - tâche envisageable : retrouver les rôles ds phrases précédentes
  - lien avec la résolution de coréférence
- Granularité variable
  - utiliser les domaines notionnels et les relations entre frames
  - pour dériver divers niveaux de granularité
  - liens avec rôles plus généraux (VerbNet)