# Annotating and parsing to semantic frames: deep syntax for FrameNet parsing

**Marie Candito**,
LLF, Paris Diderot University

joint work with
**Marianne Djemaa, Philippe Muller, Laure Vieu,**
**G. de Chalendar, B. Sagot, P. Amsili** (for the French FrameNet)
**C. Ribeyre, D. Seddah, G. Perrier, B. Guillaume** (Deep syntax project)
**Olivier Michalon, Alexis Nasr** (FrameNet parsing)

TLT 16, Prague
January, 23rd 2018

Laboratoire de linguistique formelle

université PARIS DIDEROT PARIS 7

# Outline

1. Background and motivations
2. Deep syntax project
3. French FrameNet project
   - FrameNet's characteristics
   - Methodology
   - Evaluation and stats
   - Feedback: What was difficult?
4. Deep syntax for FrameNet parsing

(For ease of reading:
examples in English when strong French/English parallelism)

# Background and Motivations

# From syntactic treebanks to shallow semantic representations

Long history of building **syntactic treebanks** for French:

- French Treebank (Abeillé and al. 03)
- Treebanks for other domains:
  - Sequoia treebank (medical, Europarl, frwiki)
  - user-generated content (Seddah et al., 12)
  - Question Bank (Seddah and Candito, 16)
- Oral French treebanks
  - Rhapsodie (Lacheret et al., 14)
  - Orfeo (Kahane et al., 17)

(among others)

# From syntactic treebanks to shallow semantic representations

Long history of building **syntactic treebanks** for French:

- French Treebank (Abeillé and al. 03)
- Treebanks for other domains:
  - Sequoia treebank (medical, Europarl, frwiki)
  - user-generated content (Seddah et al., 12)
  - Question Bank (Seddah and Candito, 16)
- Oral French treebanks
  - Rhapsodie (Lacheret et al., 14)
  - Orfeo (Kahane et al., 17)

  (among others)

But obvious lack of a
more **semantically-oriented annotated resource** for French

# Objectives

Take a step towards semantic representation of sentences

Bottom-up approach

Aiming at **generalizing** over the expression of eventualities in texts by abstracting away from:

# Objectives

Take a step towards semantic representation of sentences

Bottom-up approach

Aiming at **generalizing** over the expression of eventualities in texts by abstracting away from:

- **morpho-syntactic variation**
  - $\longrightarrow$ **deep syntax project**
    make the most of syntactic dependency trees,
    using (almost) formal criteria only

# Objectives

Move towards semantic representation of sentences

Bottom-up approach

Aiming at **generalizing** over the expression of eventualities in texts by abstracting away from:

- **lexical variation**
  - *A bomb* **exploded** *at 5 pm*
  - *A bomb* **blew up** *at 5 pm*
  - *A bomb* **explosion** *occurred at 5 pm*

# Objectives

Move towards semantic representation of sentences

Bottom-up approach

Aiming at generalizing over the expression of eventualities in texts by abstracting away from:

- lexical variation
  - *A bomb* **exploded** *at 5 pm*
  - *A bomb* **blew up** *at 5 pm*
  - *A bomb* **explosion** *occurred at 5 pm*
- predicate-argument relations
  - $\longrightarrow$ **semantic roles**
    = classes of predicate-argument relations
  - at various levels of generalization
- we chose the **FrameNet** framework

## Normalizing syntactic representations: “Deep” syntax

# Normalizing syntactic representations: “Deep” syntax

Joint work with G. Perrier, B. Guillaume, C. Ribeyre and D. Seddah
(Candito et al., 14; Perrier et al. 14)

**Bottom-up approach from syntactic parses**

- how to better take advantage of syntactic parses?
- (almost) without resorting to semantic disambiguation

# Normalizing syntactic representations: “Deep” syntax

Joint work with G. Perrier, B. Guillaume, C. Ribeyre and D. Seddah
(Candito et al., 14; Perrier et al. 14)

**Bottom-up approach from syntactic parses**

- how to better take advantage of syntactic parses?
- (almost) without resorting to semantic disambiguation

- $\longrightarrow$ use of formal criteria mostly
- to **complete and normalize syntactic valency** of verbs, adjectives

# Related work

Close to the **"Deep Syntactic representations" in MTT** (Melčuk, 88)

- in particular the Spanish AnCora-UPF corpus (Mille et al, 13)

More recently: **Enhanced dependencies** in UD (guidelines of UD v2)

- inspired by Stanford dependencies (de Marneffe and Manning, 08)
- for English (Schuster and Manning, 16), French (Candito et al. 17) ...

Also related to more **semantically-oriented dependency graphs**:

- Tectogrammatical layer in Prague Dependency Bank (Czech, English) (Hajič et al., 06)
- DeepBank (English) (Flickinger et al., 12)
- SemEval 2014 Shallow Semantic parsing shared task (Oepen et al., 14)

# Normalizing syntactic representations: “Deep” syntax

Three main enhancements, concerning very well known phenomena:

- distributing shared arguments
- neutralizing syntactic alternations (diathesis alternation)
- (by-passing morpho/syntactic markers)

# “Deep” syntax: recovering shared arguments

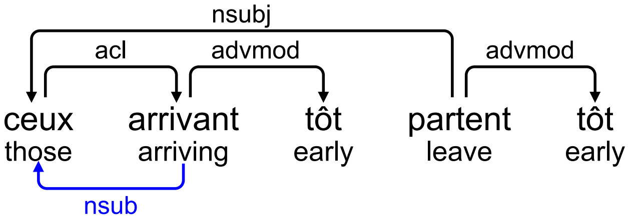
**“Subjects” of non finite verbs: cases fully determined by syntax**

- raising/control verbs: ***Paul*** *seems/wants to* ***sleep.***
- control nouns, adjectives: ***Paul****'s desire to* ***sleep.***

# “Deep” syntax: recovering shared arguments

**“Subjects” of non finite verbs: cases fully determined by syntax**

- raising/control verbs: ***Paul*** *seems/wants to* ***sleep.***
- control nouns, adjectives: ***Paul****'s desire to* ***sleep.***
- noun-modifying participles: ***those arriving*** *early /* ***arrived*** *at 9am.*
- etc...




$\longrightarrow$ dependency **graphs**

# "Deep" syntax: recovering shared arguments

**"Subjects" of non finite verbs: cases not fully determined by syntax**

- **Example: infinitive adverbial clauses**

*Il*$_{\mathbf{j}}$ *mangera avant de jouer*$_{\mathbf{j}}$
*He*$_{\mathbf{j}}$ *will-eat before to play*$_{\mathbf{j}}$
*« He will eat before playing »*

Not fully determined by syntax, but strong heuristics, e.g.:

# "Deep" syntax: recovering shared arguments

**"Subjects" of non finite verbs: cases not fully determined by syntax**

- **Example: infinitive adverbial clauses**

*Il*$_{\mathbf{j}}$ *mangera avant de jouer*$_{\mathbf{j}}$
*He*$_{\mathbf{j}}$ *will-eat before to play*$_{\mathbf{j}}$
*« He will eat before playing »*

Not fully determined by syntax, but strong heuristics, e.g.:

- When main verb is active, with non expletive subject
- $\Rightarrow$ Subject of infinitive = Subject of main verb
- in most cases (83% on Sequoia corpus)

# "Deep" syntax: recovering shared arguments

"Subjects" of non finite verbs: **cases not fully determined by syntax**

- **Example: infinitive adverbial clauses**

*Il*$_{\mathbf{j}}$ *mangera avant de jouer*$_{\mathbf{j}}$
*He*$_{\mathbf{j}}$ *will-eat before to play*$_{\mathbf{j}}$
*« He will eat before playing »*

Not fully determined by syntax, but strong heuristics, e.g.:

- When main verb is active, with non expletive subject
- $\Rightarrow$ Subject of infinitive = Subject of main verb
- in most cases (83% on Sequoia corpus)

**Counter-example:**
*D'autres photos ont subi des retouches pour **accentuer** le drame.*
*'Other photos have undergone modifications to **accentuate** the drama.'*

# “Deep” syntax: recovering shared arguments

**Example: Arguments shared by coordinated predicates**

- **Paul** *is* ***starving*** *and* ***wants*** *to* ***eat***

- **Paul** *is* ***cooking*** *and* ***selling*** *pancakes*
- *Paul is* ***cooking*** *and* ***selling*** **pancakes**

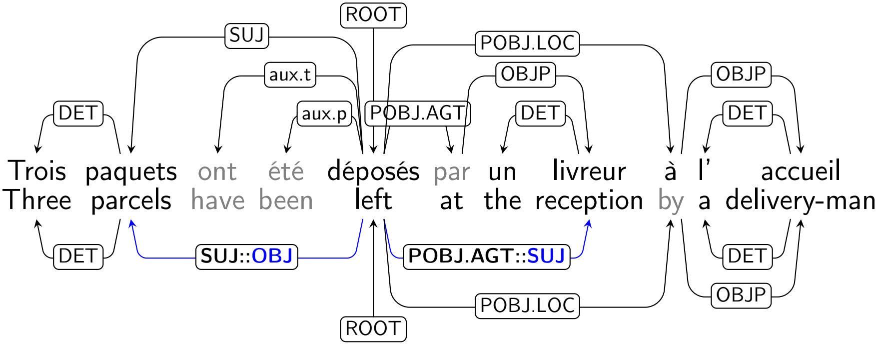
- **Paul** *is* ***sleeping*** *and* ***selling*** *pancakes*

# “Deep” syntax: Neutralizing syntactic alternations

- recover **canonical grammatical functions**
  - $\approx$ the function you would get in active personal voice
  - inspired by Relational Grammar (Perlmutter and Postal, 83)
- for French:
  - massive for passive

# "Deep" syntax: Neutralizing syntactic alternations

- recover **canonical grammatical functions**
  - ≈ the function you would get in active personal voice
  - inspired by Relational Grammar (Perlmutter and Postal, 83)
- for French:
  - massive for passive



# “Deep” syntax: Neutralizing syntactic alternations

Key choice: use canonical **syntactic** labels

- As a way to limit **argument linking diversity**
- Main diathesis alternations (passive, medipassive, causative, impersonal...)
  - have strong syntactic constraints
  - but apply either independently of or for a wide range of underlying **semantic roles**
- $\longrightarrow$ cope with these alternations at the syntactic level
  - without presupposing disambiguation of predicates, nor link to lexical entries

# "Deep" syntax: Neutralizing syntactic alternations

Key choice: use canonical syntactic labels

- As a way to limit **argument linking diversity**
- Main diathesis alternations (passive, medipassive, causative, impersonal...)
  - have strong syntactic constraints
  - but apply either independently of or for a wide range of underlying **semantic roles**
- $\longrightarrow$ cope with these alternations at the syntactic level
  - without presupposing disambiguation of predicates, nor link to lexical entries

Versus

- I, II, III ... in deep syntactic representations of MTT
- plain semantic argument numbering ARG0, ARG1, ARG2...
  (e.g. PropBank, DeepBank)
  - argument numbering refers to a semantic-aware lexicon entry
- semantic roles AGENT, PATIENT, ADDRESSEE...
  (Prague Dependency Bank tectogrammatical layer)

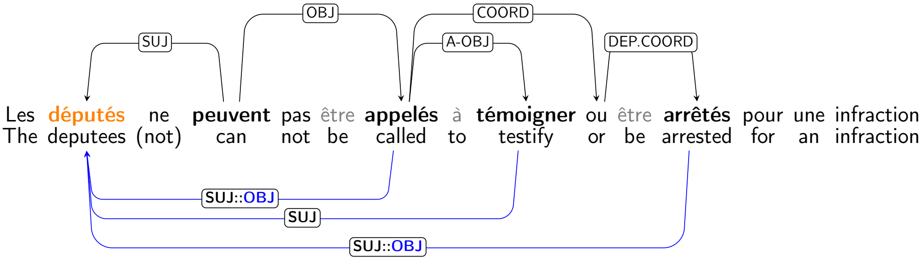
# Deep syntax: Neutralizing syntactic alternations

Handled alternations for French:

- passive
  - massive ( 18.3% non aux verbs are passives in Sequoia corpus )
  - unambiguous marking
- other alternations with morpho-syntactic marking
  - marking is in general ambiguous
  - but much rarer:
    - mediopassive ( 0.7% of non aux verbs in sequoia )
    - impersonal active ( 1% )
    - impersonal passive ( 0.27% )
    - causative ( 0.37% )
    - causative mediopassive ( absent )

# Deep syntax: interaction of phenomena

Raising, control, coordination, passive ... obvisouly interact




Note that alternation neutralization can concern deep edges!

# Deep syntax: obtaining deep syntactic graphs

**Gold data**: manually validated annotation on Sequoia (3099 sentences)

- manual annotation of 250 sentences
- to help development of deterministic graph-rewriting rules
  - Grew system (Guillaume et al. 2012)
  - OGRE system (Ribeyre et al. 2012)
- application on full Sequoia corpus (3099 sentences)
- adjudication of conflicts between the two systems
- manual checking of all infinitives and coordinations

- and further tuning of the graph-rewriting rules

# Deep syntax: obtaining deep syntactic graphs

**Gold data**: manually validated annotation on Sequoia (3099 sentences)

- manual annotation of 250 sentences
- to help development of deterministic graph-rewriting rules
  - Grew system (Guillaume et al. 2012)
  - OGRE system (Ribeyre et al. 2012)
- application on full Sequoia corpus (3099 sentences)
- adjudication of conflicts between the two systems
- manual checking of all infinitives and coordinations
- and further tuning of the graph-rewriting rules

**Quantitative assessment of enhancements**:

- 200 sentences set
- $\approx$ 1000 edges correspond to **arguments of verbs**
- 18.9% are ADDED edges (not present in surface dependency tree)
- 13.9% are ALT (have a "neutralized" label)
- $ADDED \cup ALT$ represent 27.7% of the 1000 argumental edges

**Pseudo-gold data**: graph-rewriting rules applied to French Treebank (Ribeyre et al, 2014)

- Evaluation on 200 sentences shows quality is quite good (Fscore=97.7)

# Deep syntax: obtaining deep syntactic graphs

**Pseudo-gold data**: graph-rewriting rules applied to French Treebank (Ribeyre et al, 2014)

- Evaluation on 200 sentences shows quality is quite good (Fscore=97.7)

**Deep syntax parsing**:

- pipeline surface parsing + deterministic rules
- or direct learning of graph parser (Ribeyre et al., 2015)

# French FrameNet

# Motivation

No large scale French data for (shallow) semantic analysis

⟶ let's build a resource:

- relying on **corpus evidence**
- **generalizing over:**
  - lexical items
  - predicate-to-argument relations (semantic roles)

- capitalizing on previous work for other languages

⟶ let's use the FrameNet framework (Baker et al. 1998; Fillmore 2007)

# FrameNet

Large scale "Instanciation" of the "Frame Semantics"
(Fillmore, 82)
for English
developed at Berkeley (Baker et al. 98, Fillmore et al. 03 ...)

- **frames** = schematic representations of speakers' knowledge of the situations underlying the meanings of LUs (Fillmore, 2007)
- a frame
  - can be evoked by lexical units ($\longrightarrow$ **triggers**)
    - e.g.: Commitment frame evoked by *promise.v*, *promise.n*, *oath.n* ...
  - whose semantic valency is describing using **frame-specific semantic role** names
    - Commitment; Speaker, Addressee, Message

# FrameNet

- Frame annotation = occurrence of a trigger evoking a frame, plus annotation of
  - **role fillers**: which portions of texts fill which semantic roles
  - syntactic information on role fillers

*John* ***PROMISED*** *Bill that the rights would be respected*

$\longrightarrow$ allowing to extract syntactico-semantic valency patterns

= **linking patterns**

# FrameNet: characteristics

## Structure: frames et relations entre frames

- $\approx$ 1200 frames
- linked by frame-to-frame relations (inheritance, perspective...)
- each frame-to-frame relation instance specifies role-to-role mappings
  - Complaining inherits Statement
  - Speaker $\leftrightarrow$ Complainer, Message $\leftrightarrow$ Complaint, Addressee $\leftrightarrow$ Addressee, ...

## English lexicon

- $\approx$ 13600 senses : lemma+pos / frame pairs

# FrameNet: characteristics

## Annotations

- lexicographic : examples from BNC, chosen in order to
  - capture variety of syntactic realization of semantic roles
  - and derive **linking generalizations**
  - 174000 annotated frame instances

- full-text : complete annotation of running text
  - 28000 annotated frame instances only
  - but much better training data (Das et al., 10))
  - because of natural distribution of senses and linking patterns

# FrameNet's key characteristics: Variable granularity of roles

- Well-known difficulty of defining a fixed and limited set of roles
- FrameNet's answer = **Frame-specific semantic roles**
  - $\oplus$ But coarser granularity can be derived from role-to-role mappings in frame-to-frame relation instances

# FrameNet's key characteristics: "semantics first" philosophy

Empirically grounded study of **syntax/semantic interface** is one of FrameNet's core objectives

But FrameNet's documentation (Rupenhoffer et al. 05) uses primarily **semantic criteria** for defining

- the perimeter of a frame (which generalizes over several lexemes)
- the set of roles of a frame ($\neq$Levin's classes, VerbNet...)
  - e.g. "Core" roles are those "necessary to the central meaning of the frame" (Fillmore, 2007)

# FrameNet's key characteristics: "semantics first" philosophy

- ⊖ lack of formal (syntactic) criteria to define the perimeter of frames
- ⊖ semantic granularity of the frames is partially arbitrary
  - certain frames contain synonyms only,
  - others allow more semantic variation

# FrameNet's key characteristics: "semantics first" philosophy

- $\ominus$ lack of formal (syntactic) criteria to define the perimeter of frames
- $\ominus$ semantic granularity of the frames is partially arbitrary
  - certain frames contain synonyms only,
  - others allow more semantic variation

- $\oplus$ semantic treatment of syntax/semantic mismatches ($\neq$ VerbNet)
- $\oplus$ a frame can generalize over LUs with different POS
  - e.g. the CAUSATION frame can be evoked by
  - *because.c*, *for.c*,
  - *cause.v*, *result.v*, ...
  - *cause.n*, *consequence.n*,
  - *because of.prep*, ...

# FrameNet's key characteristics: "semantics first" philosophy

- $\oplus$ roles less dependent on syntax, cf. **converse verbs** example
  - COMMERCE_SELL : sell.v, sale.n, auction.n,...
  - COMMERCE_BUY : buy.v, purchase.v, purchase.n ...
    - linked by "perspective" relation, and mapped roles

# FrameNet's key characteristics: "semantics first" philosophy

- $\oplus$ roles less dependent on syntax, cf. converse verbs example
  - COMMERCE_SELL : sell.v, sale.n, auction.n,...
  - COMMERCE_BUY : buy.v, purchase.v, purchase.n ...
    - linked by "perspective" relation, and mapped roles

- FrameNet / Propbank(Nombank) / VerbNet roles:
  - Buyer and Seller are stable in FrameNet
  - contrary to PropBank or VerbNet

I $_{\textbf{Buyer(Arg0/Agent)}}$ **bought** a Sovereign guitar $_{\textbf{Goods(Arg1/Theme)}}$
for 20 pounds $_{\textbf{Money(Arg3/Asset)}}$ from an absolute prat $_{\textbf{Seller(Arg2/Recipient)}}$

# FrameNet's key characteristics: "semantics first" philosophy

- $\oplus$ roles less dependent on syntax, cf. converse verbs example
  - COMMERCE_SELL : sell.v, sale.n, auction.n,...
  - COMMERCE_BUY : buy.v, purchase.v, purchase.n ...
    - linked by "perspective" relation, and mapped roles

- FrameNet / Propbank(Nombank) / VerbNet roles:
  - Buyer and Seller are stable in FrameNet
  - contrary to PropBank or VerbNet

I $_{\text{Buyer(Arg0/Agent)}}$ **bought** a Sovereign guitar $_{\text{Goods(Arg1/Theme)}}$ for 20 pounds $_{\text{Money(Arg3/Asset)}}$ from an absolute prat $_{\text{Seller(Arg2/Recipient)}}$

China canceled its $_{\text{Seller(Arg0/Agent)}}$ **sale** of a uranium conversion facility $_{\text{Goods(Arg1/Theme)}}$ to Iran $_{\text{Buyer(Arg2/Recipient)}}$

# ASFALDA French FrameNet project

ANR funding, Oct 2012 to March 2016
Partners: Alpage, CEA-List, LIF, LLF, MELODI (IRIT)

The objective was to build

- a French FrameNet
- and a FrameNet semantic parser

# Methodology: coverage issues

Full coverage (of course) not reachable...

Which **frames** to work on? Which **lemmas**? Which **corpus** to annotate?

# Methodology: coverage issues

Full coverage (of course) not reachable...

Which **frames** to work on? Which **lemmas**? Which **corpus** to annotate?

Preliminary full-text annotation experiments, using English frames
Feedback: **extremely difficult** to understand the perimeter of a frame

$\longrightarrow$ Decision to focus on a few **notional domains**, but fully described

# Methodology: coverage issues

- **Commercial transactions**
- **Cognitive stances**
  - Position of a Cognizer concerning the truth value of a proposition
  - stative: who knows/believes what
  - inchoative (discover.v) / causative (convince.v)
- **Verbal communication**
- **Causality**

# Methodology: overview

- Frame structure and lexicon built in parallel
  - addition of syntactic criteria to better delimitate frames
- Corpus annotation
  - Pilot annotation to build the annotation guidelines
  - 4 domain experts + 6 annotators
  - 2 independent annotations + adjudication
  - $\longrightarrow$ again some further modifications of frames and lexicon!
- Extraction of data-driven lexicon

# Frames for French

- 105 frames for French built from English frames (FN 1.5)
  - 47 non modified English frames
  - 38 adapted to French / new frame delimitation criteria
  - 20 new frames created in order to complete a domain

| Domain | Nb frames | NB LUs (with at least one annotation) |
|---|---|---|
| **Commercial Transactions** | 19 | 99 |
| **Cognitive stances** | 44 | 442 |
| **Verbal Communication** | 47 | 411 |
| **Causality** | 11 | 285 |
| **All** | 105 | 873 |

- Note: some frames belong to several domains (e.g. FR_Attempt_suasion)

**Previous coverage strategies:**

- Exhaustive frame (senses) coverage for a given lemma
  - SALSA German FrameNet (Burchardt et al., 2006)
  - produces WSD-oriented data
- versus Exhaustive lexical coverage for a given frame
  - Berkeley FrameNet
  - information extraction oriented
  - $\longrightarrow$ **chosen in French FN, for the 4 selected domains**

**Possible coverage strategies:**

- Either select examples to maximize diversity of grammatical realizations
  - lexicographic approach
  - lexicographic examples of Berkeley FrameNet

# Corpus annotation

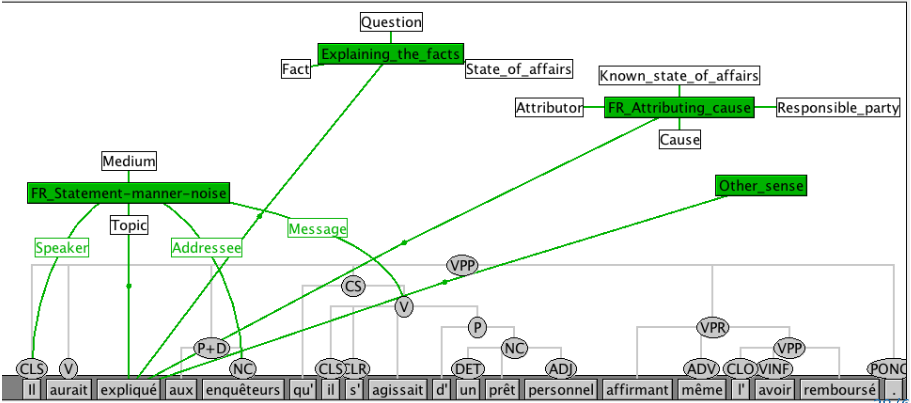
**Possible coverage strategies:**

- Either select examples to maximize diversity of grammatical realizations
  - lexicographic approach
  - lexicographic examples of Berkeley FrameNet
- or annotate running text: all lemmas of a text / first *xxx* occurrences of given lemma
  - preserves natural distribution
    - of senses
    - of grammatical characteristics of role fillers
  - used in full-text annotations of Berkeley FrameNet
  - SALSA German FrameNet
  - $\longrightarrow$ **French FN, first 100 occurrences in target corpus**

# Corpus annotation

- Annotation on syntactic (dependency) trees, in order to:
  - speed up annotation (but bias)
  - extract syntactico-semantic patterns
- 2 preexisting treebanks:
  - French Treebank (Abeillé et al. 03) and Sequoia (Candito & Seddah, 12)
  - $\approx$ 21500 sentences
  - Mainly news, plus medical, Europarl, Fr-wikipedia narrative texts
- Annotation at most first 100 occurrences of covered lemma+pos
- "Out_of_domain" dummy frame when sense is not covered

- Pilot annotation to develop annotation guide
- 2 independent annotations + adjudication (for 75% of data)
- 1 single annotation (by expert) for remaining 25%
  - Automatic pre-annotation of possible frames of the lemma
  - Salto tool (Burchardt et al., 2006)



# Evaluation: inter-annotator agreement

Between 2 independent annotations (75% of all annotations):

- for a trigger occurrence: Fscore for the frame choice
- for a frame chosen by both annotators: Fscore for role fillers

| | Nb trigger occurrences | % of Nouns | % of Verbs | Fscore inter-annotateur | | |
|---|---|---|---|---|---|---|
| | | | | Frame | Exact Role | Partial Role |
| | 17667 | 36 | 50 | 85.9 | 77.2 | 81.9 |
| Break-down by domain | | | | | | |
| Commercial | 3307 | 60 | 40 | 92.0 | 73.4 | 80.4 |
| Causality | 7691 | 30 | 48 | 79.2 | 74.2 | 80.4 |
| Cognitive stances | 7886 | 28 | 62 | 90.6 | 81.1 | 86.0 |
| Communication | 2221 | 23 | 76 | 89.6 | 82.3 | 87.5 |
| Break-down by trigger POS | | | | | | |
| V | 8834 | - | - | 87.6 | 82.8 | 87.1 |
| N | 6234 | - | - | 86.8 | 68.3 | 72.5 |
| other | 2509 | - | - | 77.7 | 74.6 | 82.1 |

# Stats (1.2 release)

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

- $\longrightarrow \approx$ 16200 annotated frames (plus 8750 occurrences "Other_sense")

| | Nb distinct frames | Nb distinct lemma+POS | Nb senses | Nb annotated frames ($\neq$ Other_sense) |
|---|---|---|---|---|
| ALL | 105 | 873 | 1109 | **16167** |
| Fully covered lemma+POS | | 490 | | **7213** |
| Commercial | 19 | 90 | 99 | 2930 |
| Causality | 11 | 243 | 285 | 3895 |
| Cognitive | 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 |

# Semantic roles

- 421 frame-specific roles (for the 105 frames)
- grouped into 40 "macro-roles" defined at the domain level

See/get data at `http://asfalda.linguist.univ-paris-diderot.fr/frameIndex.xml`

# Frame structure and lexicon: feedback

- Division into frames is definitely the most difficult part
  - When starting from English frames: difficulty to understand differences between frames
  - Attempt to specify distinctive characteristics of frames

- $\longrightarrow$ Automatic induction ?
  - e.g. (Modi et al., 12)
  - with or without manual intervention

# Corpus FrameNet annotation: typical problems

Berkeley FrameNet: **chosen** examples $\longrightarrow$ Annotation difficulties not much documented

Very wide range of phenomena to handle

- divergences between litteral expression and interpretation
  - multi-word expressions (for triggers)
  - metaphors, ellipsis...
- lexical semantics
  - polysemy
  - nouns referring to a participant

# Corpus FrameNet annotation: typical problems

Berkeley FrameNet: **chosen** examples $\longrightarrow$ Annotation difficulties not much documented

Very wide range of phenomena to handle

- divergences between litteral expression and interpretation
  - multi-word expressions (for triggers)
  - metaphors, ellipsis...
- lexical semantics
  - polysemy
  - nouns referring to a participant
- morpho-syntax / semantics interface
  - syntax/semantic mismatches
  - elliptic coordinations: arg cluster, head gapping...
  - ambiguous status of reflexive "se" clitic
    - diathesis alternation marker (mediopassive, neuter)
    - frozen
    - true reflexive ...

# Focus: predicative noun referring to a participant

(English examples for ease of reading)

**Relational nouns:**

*The precise **CAUSES** of multiple sclerosis are not yet known.*

**SYNTACTICALLY:**

- *causes* is **monovalent**: one PP

# Focus: predicative noun referring to a participant

(English examples for ease of reading)

**Relational nouns:**

*The precise **CAUSES** of multiple sclerosis are not yet known.*

**SYNTACTICALLY:**

- *causes* is **monovalent**: one PP

**SEMANTICALLY:** predicate could be

- **monovalent**: **entity_being_the_cause_of(Effect)**

# Focus: predicative noun referring to a participant

(English examples for ease of reading)

**Relational nouns:**

*The precise* ***CAUSES*** *of multiple sclerosis are not yet known.*

**SYNTACTICALLY:**

- *causes* is **monovalent**: one PP

**SEMANTICALLY:** predicate could be

- **monovalent**: **entity_being_the_cause_of(Effect)**
- $\longrightarrow$ **bivalent**: **causality_relation(Cause, Effect)**

# Focus: predicative noun referring to a participant

(English examples for ease of reading)

**Relational nouns:**

*The precise **CAUSES** of multiple sclerosis are not yet known.*

**SYNTACTICALLY:**

- *causes* is **monovalent**: one PP

**SEMANTICALLY:** predicate could be

- **monovalent**: **entity_being_the_cause_of(Effect)**
- $\longrightarrow$ **bivalent**: **causality_relation(Cause, Effect)**

(frame Causation; Cause, Effect)

*The precise **CAUSES** of multiple sclerosis are not yet known*

# Focus: predicative noun referring to a participant

In annotation guidelines, explicit distinction of:

- **referential uses** (most occurrences)
- **predicative uses**
  - $\longrightarrow$ in which we consider the trigger does not refer to the participant

Distinction pertaining for any noun:

- *Have you seen the unicorn?*
- *This animal is a unicorn.*

# Focus: predicative noun referring to a participant

## Referential uses

(frame Causation; Cause, Effect)

*The exact **CAUSES** of multiple sclerosis are not yet known*

*Liberia is still suffering from the **CONSEQUENCES** of 14 years of war*

# Focus: predicative noun referring to a participant

## Typical cases of predicative uses:

(frame Causation; Cause, Effect)

- **Apposition**:
  *First **CAUSE** of unemployment , short-term contract terminations are rising.*
- **Copulative sentence**:
  *Chronic venous insufficiency is the **CONSEQUENCE** of various abnormalities.*
- **Inverted copulative sentence**:
  *The **RESULT** of these measures has been to partially fill the deficit.*

# Event/Participant ambiguity

Typical case of a predicate that can refer to a role:
event/participant ambiguous nominalisations (or event/result).

(Commerce_buy; Buyer, Goods)

- *(...) to multiply its **PURCHASES** of car parts (...)*
  - event, buying act
- *How many times have you put down your **PURCHASES** willy-nilly, and trusted the cashier to do his job right?*
  - *purchases* refers to bought goods

# Deep syntax for FrameNet parsing

# FrameNet parsing

- WSD task: frame selection for an ambiguous trigger
- SRL task: role identification

Challenges

- Generalization over data
  - WordNet (e.g. Johansson et Nugues, 2007)
  - Distributed representations (e.g. Hermann et al. ACL 2014)
- Joint models (e.g. Yang and Mitchell, 2017)

# Deep syntax for FrameNet parsing

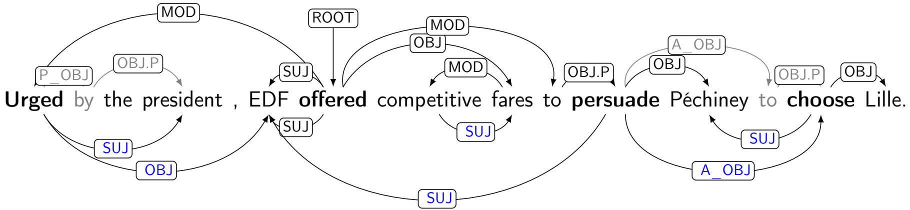
Joint work **Olivier Michalon**, Corentin Ribeyre, Alexis Nasr
(Michalon et al. Coling 2016)

- syntactic features known to be quite useful for SRL
  - since Gildea et Jurafsky, 2002
  - still true with neural networks approach
    (Hermann et al. 14; Yang and Mitchell 17)

- is it worth using deep syntax ?

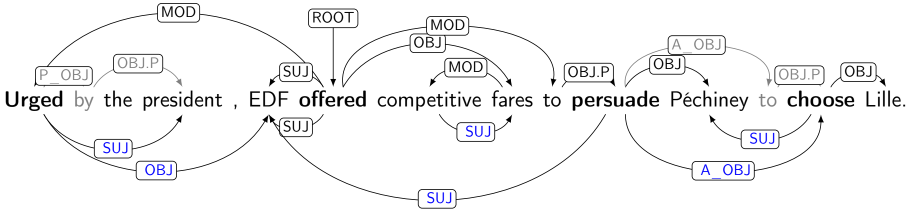
# Deep syntax for FrameNet parsing

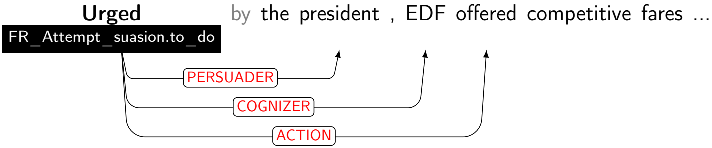
(arcs for determiners and punctuations not shown)



# Deep syntax for FrameNet parsing

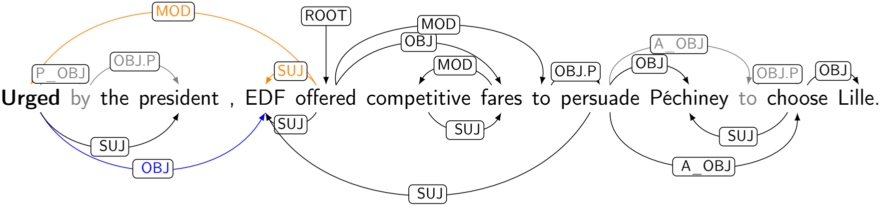
(arcs for determiners and punctuations not shown)

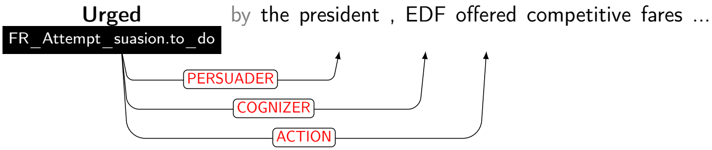


# Deep syntax for FrameNet parsing

(arcs for determiners and punctuations not shown)




Syntactic path between "Urged" et "EDF" :

- surface: -mod,+suj
- deep: +obj

# Measuring the normalizing effect

Syntactic path between

- a predicate
- (the syntactic head) of a role filler

For a given role, deep syntactic paths are **more regular**:

**The distributions P(path to role filler | frame-specific role) are less scattered when using deep syntax.**

# Measuring the normalizing effect

Syntactic path between

- a predicate
- (the syntactic head) of a role filler

For a given role, deep syntactic paths are **more regular**:

**The distributions P(path to role filler | frame-specific role) are less scattered when using deep syntax.**

Indeed average entropy over all roles decreases:

- **1.65** with "classical" syntactic paths
- **1.32** with "deep" syntactic paths

# Measuring the normalizing effect

5 most frequent paths,
for the role fillers of verbal triggers

| surface syntax | | deep syntax | |
|---|---|---|---|
| (+suj) | 25.0% | (+suj) | 33.1% |
| (+obj) | 17.0% | (+obj) | 32.8% |
| (-mod) | 8.0% | (+a_obj) | 4.7% |
| (+obj,+obj.cpl) | 4.4% | (-mod) | 3.2% |
| (+a_obj,+obj.p) | 4.1% | (+mod,+obj.p) | 2.5% |
| Total | 58.6 % | Total | 76.2 % |

# Impact for FrameNet parsing

Very basic system (pipeline WSD + SRL, supervised linear classification)

- WSD : one classifier per ambiguous lemma
- SRL : one classifier per frame

Positive impact for FrameNet SRL, in particular for verbal triggers

| | Prec. | | Recall | | F-measure | |
|---|---|---|---|---|---|---|
| Input syntax | surf | deep | surf | deep | surf | deep |
| WSD (gold frame $\neq$ 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

- Deep syntax:
  - Currently checking whether improvement carries over using a neural architecture

# Perspectives

- Deep syntax:
  - Currently checking whether improvement carries over using a neural architecture
- FrameNet construction:
  - very difficult and time consuming
  - coverage will never be sufficient
  - $\longrightarrow$ Frame induction
    - WSD + deep syntactic subcats + distributed representation of fillers...

# Perspectives

- Deep syntax:
  - Currently checking whether improvement carries over using a neural architecture
- FrameNet construction:
  - very difficult and time consuming
  - coverage will never be sufficient
  - $\longrightarrow$ Frame induction
    - WSD + deep syntactic subcats + distributed representation of fillers...
- Saturation of annotated frames
  - a lot of roles are not instantiated within sentence
  - recover role fillers in preceding sentences

# Perspectives

- Deep syntax:
  - Currently checking whether improvement carries over using a neural architecture
- FrameNet construction:
  - very difficult and time consuming
  - coverage will never be sufficient
  - $\longrightarrow$ Frame induction
    - WSD + deep syntactic subcats + distributed representation of fillers...
- Saturation of annotated frames
  - a lot of roles are not instantiated within sentence
  - recover role fillers in preceding sentences
- In the long term, still a long way for NLU
  - in particular: factivity of eventualities

Thank you!

# Bibliography I


Anne ABEILLE, Lionel CLEMENT et François TOUSSENEL : Building a treebank for french. In Treebanks : Building and Using Parsed Corpora, pages 165–188. Springer, 2003.

Collin F. BAKER, Charles J. FILLMORE et John B. LOWE : The berkeley framenet project. ACL '98, pages 86–90, Stroudsburg, PA, USA, 1998.

Hans C. BOAS : Semantic frames as interlingual representations for multilingual lexical databases. In Hans Christian BOAS, ed. : Multilingual FrameNets in computational lexicography : methods and applications. Trends in linguistics. Mouton de Gruyter, Berlin, New York, 2009.

Aljoscha BURCHARDT, Katrin ERK, Anette FRANK, Andrea KOWALSKI et Sebastian PADO : SALTO : A versatile multi-level annotation tool. LREC 2006, Genoa, Italy, 2006.

Marie CANDITO , Bruno GUILLAUME, Guy PERRIER and Djamé SEDDAH : Enhanced UD Dependencies with Neutralized Diathesis Alternation, DEPLING 2017, Pisa, Italy, 2017.

Marie CANDITO, Guy PERRIER, Bruno GUILLAUME, Corentin RIBEYRE, Karën FORT, Djamé SEDDAH et Eric VILLEMONTE DE LA CLERGERIE : Deep syntax annotation of the sequoia french treebank. LREC, Reykjavik, Iceland, 2014.

Marie CANDITO et Djamé SEDDAH : Le corpus sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. TALN'2012, pages 321–334, Grenoble, France, 2012.

Marianne DJEMAA, Marie CANDITO, Philippe MULLER and Laure VIEU : Corpus annotation within the French Framenet: methodology and results, LREC 2016, Portoroz, Slovenia, 2016.

# Bibliography II


Charles J. FILLMORE : Valency issues in framenet. In Thomas HERBST et Katrin GOTZ-VOTTELER, eds : Valency : Theoretical, descriptive and cognitive issues, volume 187 de Trends in Linguistics. Studies and Monographs, pages 129–160. Walter de Gruyter, 2007.

Dan FLICKINGER, Yi ZHANG and Valia KORDONI : DeepBank. A dynamically annotated treebank of the Wall Street Journal. TLT'11, Lisbon, Portugal, 2012.

Jan HAJIČ, Jarmila PANEVOVÁ, Eva HAJICOVÁ, Petr SGALL, Petr PAJAS, Jan ŠTEPÁNEK, Jiří HAVELKA, Marie MIKULOVÁ, M., Zdenek ZABOKRTSKỲ, and Magda Š. Razımová : Prague dependency treebank 2.0. CD-ROM, Linguistic Data Consortium, LDC Catalog No.: LDC2006T01, Philadelphia, 2006.

Karl Moritz HERMANN, Dipanjan DAS, Jason WESTON, and Kuzman GANCHEV : Semantic frame identification with distributed word representations. ACL 2014, Baltimore, USA, 2014.

Bruno GUILLAUME, Guillaume BONFANTE, MASSON, Mathieu MOREY, and Guy PERRIER : Grew : un outil de réécriture de graphes pour le TAL. TALN 2012, Grenoble, France, 2012.

Laura KALLMEYER and Rainer OSSWALD : Syntax-Driven Semantic Frame Composition in Lexicalized Tree Adjoining Grammars. Journal of Language Modelling 1(2):267-330, 2013.

Karin KIPPER, Anna KORHONEN, Neville RYANT et Martha PALMER : A large-scale classification of english verbs. Language Resources and Evaluation, 42(1):21–40, 2008.

Anne LACHERET, Sylvain KAHANE, Julie BELIAO, Anne DISTER, Kim GERDES, Jean-Philippe GOLDMAN, Nicolas OBIN, Paola PIETRANDREA and Atanas TCHOBANOV :

# Bibliography III


Rhapsodie: un Treebank annoté pour l'étude de l'interface syntaxe-prosodie en français parlé. CMLF 2014, Berlin Germany, 2014.

Igor MEL'ČUK : Dependency Syntax: Theory and Practice. The SUNY Press, Albany, NY, 1988.

Olivier MICHALON, Corentin RIBEYRE, Marie CANDITO et Alexis NASR : Deeper syntax for better semantic parsing. Coling 2016, Osaka, Japan, 2016.

Simon MILLE, Alicia BURGA and Léo WANNER : AnCoraUPF: A Multi-Level Annotation of Spanish. DEPLING 2013, Prague, Czech Republic, 2013.

Ashutosh MODI, Ivan TITOV and Alexandre KLEMENTIEV : Unsupervised Induction of Frame-Semantic Representations. NAACL-HLT Workshop on the Induction of Linguistic Structure, Montréal, Canada, 2012.

Stephan OEPEN, Marco KUHLMANN, Yusuke MIYAO, Daniel ZEMAN, Dan FLICKINGER, Jan HAJIČ, Angelina IVANOVA, and Yi ZHANG. Semeval 2014 task 8: Broad-coverage semantic dependency parsing. SemEval 2014, Dublin, Ireland, 2014.

Martha PALMER, Daniel GILDEA et Paul KINGSBURY : The proposition bank : An annotated corpus of semantic roles. Computational linguistics, 31(1):71–106, 2005.

Daniel PERLMUTTER and Peter POSTAL (eds.) : Studies in Relational Grammar 1. University of Chicago Press, Chicago, 1983.

Corentin RIBEYRE, Marie CANDITO et Djamé SEDDAH : Semi-automatic deep syntactic annotations of the french treebank. TLT 13, Tubingen, Germany, 2014.

# Bibliography IV


Sebastian SCHUSTER and Christopher D. MANNING : Enhanced English Universal Dependencies: An Improved Representation for Natural Language Understanding Tasks. LREC 2016, Portoroz, Slovenia, 2016.

Djamé SEDDAH, Benoit SAGOT, Marie CANDITO, Virginie MOUILLERON and COMBET Vanessa : The French Social Media Bank: a Treebank of Noisy User Generated Content, COLING 2012, Mumbay, India, 2012.

Djamé SEDDAH and Marie CANDITO : Hard Time Parsing Questions: Building a QuestionBank for French Djamé Seddah, LREC 2016, Portoroz, Slovenia, 2016.

Josef RUPPENHOFER, Michael ELLSWORTH, Miriam R.L. PETRUCK, Christopher R. JOHNSON et Jan SCHEFFCZYK : FrameNet II : Extended Theory and Practice. International Computer Science Institute, Berkeley, California, 2006. Distributed with the FrameNet data.

Bishan YANG and Tom Mitchell : A Joint Sequential and Relational Model for Frame-Semantic Parsing. EMNLP 2017, Copenhagen, Denmark, 2017.