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
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)
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
  - 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**
  - → 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
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

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

\[
\begin{array}{c}
\text{ceux} \\
\text{arrivant} \\
\text{tôt} \\
\text{partent} \\
\text{tôt}
\end{array}
\]

\[
\begin{array}{c}
\text{those} \\
\text{arriving} \\
\text{early} \\
\text{leave} \\
\text{early}
\end{array}
\]

\(\rightarrow\) dependency graphs

11/62
“Deep” syntax: recovering shared arguments

“Subjects” of non finite verbs: cases not fully determined by syntax

- Example: infinitive adverbial clauses

Il mangerá avant de jouer
He will-eat before to play
« 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 j\textsubscript{e} manger\textit{a} avant de jouer\textit{j}
He\textsubscript{j} will-\textit{e}at before to play\textit{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 mangera avant de jouer
He will-eat before to play
« He will eat before playing »

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

- When main verb is active, with non expletive subject
- 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.’
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**
  - ≈ the function you would get in active personal voice
  - inspired by Relational Grammar (Perlmutter and Postal, 83)
- for French:
  - massive for passive

Trois paquets ont été déposés par un livreur à l’accueil.

Three parcels have been left at the reception by a delivery-man.
“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

Trois paquets ont été déposés par un livreur à l’accueil
Three parcels have been left at the reception by a delivery-man
Root requires 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
- 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**
-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)
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)
Raising, control, coordination, passive ... obviously interact

Les députés ne peuvent pas être appelés à témoigner ou être arrêtés pour une infraction

The deputees (not) can not be called to testify or be arrested for an infraction

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

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
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)
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 (→ 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

→ allowing to extract syntactico-semantic valency patterns
= linking patterns
FrameNet: characteristics

Structure: frames et relations entre frames

• ≈ 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 ↔ Complainer, Message ↔ Complaint, Addressee ↔ Addressee, ...

English lexicon

• ≈ 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**
  - But coarser granularity can be derived from role-to-role mappings in frame-to-frame relation instances
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 (≠ 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

- $\Theta$ lack of formal (syntactic) criteria to define the perimeter of frames
- $\Theta$ 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

- ⊖ 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

- ⊕ semantic treatment of syntax/semantic mismatches (≠ VerbNet)
- ⊕ 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

• ⊕ 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

• ⊕ 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 Buyer(Arg0/Agent) bought a Sovereign guitar Goods(Arg1/Theme)
for 20 pounds Money(Arg3/Asset) from an absolute prat Seller(Arg2/Recipient)
FrameNet’s key characteristics: "semantics first" philosophy

- ⊕ 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  Buyer(Arg0/Agent) bought a Sovereign guitar Goods(Arg1/Theme)
for 20 pounds Money(Arg3/Asset) from an absolute prat Seller(Arg2/Recipient)

China canceled its Seller(Arg0/Agent) sale of a uranium conversion facility Goods(Arg1/Theme) to Iran Buyer(Arg2/Recipient)
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?
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

→ 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
  - $\longleftarrow$ 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

<table>
<thead>
<tr>
<th>Domain</th>
<th>Nb frames</th>
<th>NB LUs (with at least one annotation)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Commercial Transactions</td>
<td>19</td>
<td>99</td>
</tr>
<tr>
<td>Cognitive stances</td>
<td>44</td>
<td>442</td>
</tr>
<tr>
<td>Verbal Communication</td>
<td>47</td>
<td>411</td>
</tr>
<tr>
<td>Causality</td>
<td>11</td>
<td>285</td>
</tr>
<tr>
<td>All</td>
<td>105</td>
<td>873</td>
</tr>
</tbody>
</table>

- 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
  - → chosen in French FN, for the 4 selected domains
Corpus annotation

Possible coverage strategies:

- Either select examples to maximize diversity of grammatical realizations
  - lexicographic approach
  - lexicographic examples of Berkeley FrameNet
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 \( \text{xxx} \) 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
  - \( \rightarrow \) 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)
  - ≈ 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
Corpus annotation: Graphical tool

- 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

<table>
<thead>
<tr>
<th></th>
<th>Nb trigger occurrences</th>
<th>% of Nouns</th>
<th>% of Verbs</th>
<th>Fscore inter-annotateur</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Frame</td>
</tr>
<tr>
<td>Total</td>
<td>17667</td>
<td>36</td>
<td>50</td>
<td>85.9</td>
</tr>
</tbody>
</table>

Break-down by domain

<table>
<thead>
<tr>
<th>Domain</th>
<th>Nb trigger occurrences</th>
<th>% of Nouns</th>
<th>% of Verbs</th>
<th>Frame</th>
<th>Exact Role</th>
<th>Partial Role</th>
</tr>
</thead>
<tbody>
<tr>
<td>Commercial</td>
<td>3307</td>
<td>60</td>
<td>40</td>
<td>92.0</td>
<td>73.4</td>
<td>80.4</td>
</tr>
<tr>
<td>Causality</td>
<td>7691</td>
<td>30</td>
<td>48</td>
<td>79.2</td>
<td>74.2</td>
<td>80.4</td>
</tr>
<tr>
<td>Cognitive stances</td>
<td>7886</td>
<td>28</td>
<td>62</td>
<td>90.6</td>
<td>81.1</td>
<td>86.0</td>
</tr>
<tr>
<td>Communication</td>
<td>2221</td>
<td>23</td>
<td>76</td>
<td>89.6</td>
<td>82.3</td>
<td>87.5</td>
</tr>
</tbody>
</table>

Break-down by trigger POS

<table>
<thead>
<tr>
<th>POS</th>
<th>Nb trigger occurrences</th>
<th>% of Nouns</th>
<th>Fscore inter-annotateur</th>
</tr>
</thead>
<tbody>
<tr>
<td>V</td>
<td>8834</td>
<td>-</td>
<td>87.6</td>
</tr>
<tr>
<td>N</td>
<td>6234</td>
<td>-</td>
<td>86.8</td>
</tr>
<tr>
<td>other</td>
<td>2509</td>
<td>-</td>
<td>77.7</td>
</tr>
</tbody>
</table>
Stats (1.2 release)

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

- ≈ 16200 annotated frames (plus 8750 occurrences “Other_sense”)

<table>
<thead>
<tr>
<th></th>
<th>Nb distinct frames</th>
<th>Nb distinct lemma+POS</th>
<th>Nb senses</th>
<th>Nb annotated frames (≠ Other_sense)</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALL</td>
<td>105</td>
<td>873</td>
<td>1109</td>
<td>16167</td>
</tr>
<tr>
<td>Fully covered</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>lemma+POS</td>
<td>490</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Commercial</td>
<td>19</td>
<td>90</td>
<td>99</td>
<td>2930</td>
</tr>
<tr>
<td>Causality</td>
<td>11</td>
<td>243</td>
<td>285</td>
<td>3895</td>
</tr>
<tr>
<td>Cognitive</td>
<td>44</td>
<td>372</td>
<td>442</td>
<td>5426</td>
</tr>
<tr>
<td>Communication</td>
<td>47</td>
<td>347</td>
<td>411</td>
<td>5233</td>
</tr>
<tr>
<td>N</td>
<td>-</td>
<td>296</td>
<td>346</td>
<td>5282</td>
</tr>
<tr>
<td>V</td>
<td>-</td>
<td>446</td>
<td>594</td>
<td>9165</td>
</tr>
<tr>
<td>PREP</td>
<td>-</td>
<td>35</td>
<td>43</td>
<td>674</td>
</tr>
<tr>
<td>ADV</td>
<td>-</td>
<td>26</td>
<td>42</td>
<td>407</td>
</tr>
<tr>
<td>CONJ</td>
<td>-</td>
<td>22</td>
<td>28</td>
<td>301</td>
</tr>
<tr>
<td>ADJ</td>
<td>-</td>
<td>43</td>
<td>48</td>
<td>234</td>
</tr>
</tbody>
</table>
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
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

→ Automatic induction?
  - e.g. (Modi et al., 12)
  - with or without manual intervention
Corpus FrameNet annotation: typical problems

Berkeley FrameNet: chosen examples \(\rightarrow\) Annotation difficulties not much documented

Very wide range of phenomena to handle

- divergences between literal 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 $\rightarrow$ Annotation difficulties not much documented

Very wide range of phenomena to handle

- divergences between literal 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 ...
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)`
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)
- → **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)
- → **bivalent**: causality_relation(Cause, Effect)

(frame Causation; Cause, Effect)

*The precise CAUSES of multiple sclerosis are not yet known*
In annotation guidelines, explicit distinction of:

- **referential uses** (most occurrences)
- **predicative uses**
  - → 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.*
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
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)

- \((...) \text{ to multiply } \text{its PURCHASES of car parts } (...)\)
  - event, buying act

- \text{How many times have you put down } \text{your PURCHASES willy-nilly, and trusted the cashier to do his job right?}
  - \(\text{purchases}\) 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)

Urged by the president, EDF offered competitive fares to persuade Péchiney to choose Lille.
Deep syntax for FrameNet parsing

Urged by the president, EDF offered competitive fares to persuade Péciney to choose Lille.
Deep syntax for FrameNet parsing

(arcs for determiners and punctuations not shown)

Urged by the president, EDF offered competitive fares to persuade Péchiney to choose Lille.

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 $P(\text{path to role filler} \mid \text{frame-specific role})$ 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 $P(\text{path to role filler} \mid \text{frame-specific role})$ 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

<table>
<thead>
<tr>
<th>surface syntax</th>
<th>deep syntax</th>
</tr>
</thead>
<tbody>
<tr>
<td>(+suj)</td>
<td>(+suj)</td>
</tr>
<tr>
<td>25.0%</td>
<td>33.1%</td>
</tr>
<tr>
<td>(+obj)</td>
<td>(+obj)</td>
</tr>
<tr>
<td>17.0%</td>
<td>32.8%</td>
</tr>
<tr>
<td>(-mod)</td>
<td>(+a_obj)</td>
</tr>
<tr>
<td>8.0%</td>
<td>4.7%</td>
</tr>
<tr>
<td>(+obj,+obj.cpl)</td>
<td>(-mod)</td>
</tr>
<tr>
<td>4.4%</td>
<td>3.2%</td>
</tr>
<tr>
<td>(+a_obj,+obj.p)</td>
<td>(+mod,+obj.p)</td>
</tr>
<tr>
<td>4.1%</td>
<td>2.5%</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>Total</strong></td>
</tr>
<tr>
<td>58.6 %</td>
<td>76.2 %</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Input syntax</th>
<th>Prec.</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>surf</td>
<td>80.1</td>
<td>80.7</td>
<td>80.1</td>
</tr>
<tr>
<td>deep</td>
<td>80.1</td>
<td>80.7</td>
<td>80.1</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>WSD (gold frame ≠ Other_sense)</th>
<th>Prec.</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>surf</td>
<td>81.4</td>
<td>86.4</td>
<td>59.1</td>
</tr>
<tr>
<td>deep</td>
<td>66.1</td>
<td>68.5</td>
<td>68.5</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SRL (for gold role filler heads)</th>
<th>Prec.</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td>surf</td>
<td>75.7</td>
<td>80.3</td>
<td>51.6</td>
</tr>
<tr>
<td>deep</td>
<td>61.3</td>
<td>68.0</td>
<td>68.0</td>
</tr>
</tbody>
</table>

Table: FastSem results for verbs, using gold (top) and predicted (bottom) surf and deep syntax.
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
  - → 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
• Deep syntax:
  ▶ Currently checking whether improvement carries over using a neural architecture

• FrameNet construction:
  ▶ very difficult and time consuming
  ▶ coverage will never be sufficient
  ▶ 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
  - \( \rightarrow \) 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!


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


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


