Developing a French FrameNet: Methodology and First Results


* Alpage (Univ. Paris Diderot / INRIA), LLF (Univ. Paris Diderot / CNRS), † Univ. Paris 13, ○ IRIT, Toulouse University, Δ IRIT, CNRS, ◇ CEA LIST, ¶ Lattice (Univ. Paris 3, CNRS, ENS), ★ CLILLAC-ARP (Univ. Paris Diderot)
asfalsa@inria.fr

Background and Motivation

ASFALDA project (ANR-CORD-023)
- a 3-year ANR project (started in sept. 2012)
- → build FrameNet-annotated resources for French
- → build FrameNet-analyzer

Motivations

Why FrameNet?
- → semantic-oriented generalizations, initially on English
- → proved to be portable across languages (Boas et al., 05)

Why French?
- → French lacks large-scale semantically annotated resources

Annotation strategy

Focus on a set of domains
- We build on the English FrameNet
- Frame definition, frame/lemma associations and corpus annotations are very time-consuming tasks
- Same coverage not reachable → we focus on a set of domains
  - → work domain-by-domain
  - (although domains do overlap)
  - → enforce coherence and exhaustivity within a domain

Notional domains

Linguistic communication
- e.g. dire (say), prétendre (pretend) ...
- pervasive in our corpora
- representation challenge: quotation with non-communicative verbs (Sagot and Danlos, 2010)
  - Example: “It would be wrong to neglect ... India”, the prefect corrected.
  - potential information extraction: who said what to whom when?

Cognitive positions
- e.g. penser (think), savoir (know), s’apercevoir (realize) ...
- varying degree of certainty on a belief
- potentially useful for fact extraction

Judgment/Evaluation
- e.g. critique (critique), apprécier (appreciate), beau (beautiful) ...
- potential application: opinion mining

“Cross-Cutting Domains”

Temporal relations / Spatial relations / Causality
- pervasive semantic phenomena
- at the interface between lexical semantics and discourse semantics
- evoked by verbs, deverbals, but also prepositions, conjunctions, adverbs
- captured by frames and frame-to-frame relations

Current status

Frames and lexicon for our target domains are now fully defined. Work on frames and lexicon was performed jointly and iteratively:
- → starting from English frames
- → using a preliminary French framenet lexicon automatically projected from English (Paldi, 07; Mouton et al. 2010)

Frames

Understanding the exact semantic scope of a frame often difficult
- → we decided to:
  - create expert teams for each domain
  - systematically make explicit the distinctive characteristics of frames
  - frame merging if the distinction seemed too subtle, or not justified for French

<table>
<thead>
<tr>
<th>Notional domain</th>
<th>Nb. of frames</th>
<th>Total nb. modified frames</th>
<th>Merges</th>
<th>Splits</th>
<th>Modifications of roles only</th>
<th>Other modifications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Commerical transaction</td>
<td>11</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Verbal communication</td>
<td>34</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Judgement/Evaluation</td>
<td>16</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Cognitive positions</td>
<td>19</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Temporal relations</td>
<td>4</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Spatial relations</td>
<td>20</td>
<td>14</td>
<td>2</td>
<td>12</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Causality</td>
<td>8</td>
<td>5</td>
<td>1</td>
<td>0</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>Total</td>
<td>106</td>
<td>28</td>
<td>8</td>
<td>16</td>
<td>3</td>
<td>3</td>
</tr>
</tbody>
</table>

Lexicon

- Initial scenario: double annotation by external annotators + adjudication
- but the task proved to be very difficult!

- Actual scenario:
  - double annotation by experts + adjudication
  - additional phase of polysemy checking
  - and at each stage, frequent back-and-forths between lexicon development and frame definitions / remodeling

- Resulting lexicon: 1638 frame/lemma pairs, for 1359 distinct lemmas
  - (797 verbs, 574 nouns, 270 adjectives, 119 prepositions, 41 conjunctions)

Next Steps

Semantic annotations on treebanks

- Annotation of frames and semantic roles on top of syntactic dependency trees
- Using the Salto graphical interface (Buchardt et al., 2006)
- Using two treebanks converted to dependencies:
  - The French Treebank (Abeillé and Barrier, 2004): 18k sentences, news
  - The Sequoia Treebank (Candito and Seldath, 2012): 3k sentences, medical, europarl, news
  - → usable to study syntax-semantic interface

- temporal and location modifiers treated using time- and location-related frames instead of peripheral semantic roles

The General Secretary said during a press conference that he thought “the situation will unlock” before the deadline of February 15

Figure: Example of target annotations on dependency trees, with the Salto tool