Developing a French FrameNet: Methodology and First results

Marie Candito[⋆], Pascal Amsili[•], Lucie Barque[‡⋆], Farah Benamara[◇], Gaël de Chalendar[♯], Marianne Djemaa[⋆], Pauline Haas[‡▽], Richard Huyghe[○], Yvette Yannick Mathieu[•], Philippe Muller[◇], Benoît Sagot[⋆], Laure Vieu[△]

⋆ Alpage (Univ. Paris Diderot / INRIA), • LLF (Univ. Paris Diderot / CNRS), ‡ Univ. Paris 13,
◇ IRIT, Toulouse University , △ IRIT, CNRS, ♯ CEA LIST,
▽ Lattice (Univ. Paris 3, CNRS, ENS), ○ CLILLAC-ARP (Univ. Paris Diderot)
asfalda@inria.fr

Abstract

The Asfalda project aims to develop a French corpus with frame-based semantic annotations and automatic tools for shallow semantic analysis. We present the first part of the project: focusing on a set of notional domains, we delimited a subset of English frames, adapted them to French data when necessary, and developed the corresponding French lexicon. We believe that working domain by domain helped us to enforce the coherence of the resulting resource, and also has the advantage that, though the number of frames is limited (around a hundred), we obtain full coverage within a given domain.

1. Introduction

The ASFALDA project[1] is a three-year project which started in October 2012, with the objective of building semantic resources and a corresponding semantic analyzer for French, able to capture both generalizations over predicates and over the semantic arguments of predicates. We chose to build on the work resulting from the FrameNet project (Baker et al., 1998), hence the resources built within ASFALDA will consist in a French lexicon in which lexical units are associated to FrameNet frames, and a semantic annotation layer added on top of existing syntactic French treebanks (the French Treebank (Abeillé and Barrier, 2004) and the Sequoia treebank (Candito and Seddah, 2012)). The project also aims at investigating new models for frame-based semantic analysis.

In this paper we focus on the domain-by-domain strategy we adopted to build the French FrameNet, on the methodology for adapting frames from English to French, and on the lexicon development.

The original FrameNet project[2] provides a structured set of prototypical situations, called *frames*, along with a semantic characterization of the participants of these situations (called *frame elements*, FEs). Frame development was carried along annotation of lexicographic examples, extracted from the BNC. While other English semantic resources (such as PropBank (Palmer et al., 2005) or VerbNet (Schuler, 2005)) also provide semantic classes and/or semantic roles for predicate arguments, we chose FrameNet mainly because of its more semantic orientation, which is crucial for portability to other languages. FrameNet offers generalization both over syntactic variation (e.g. diathesis alternation) but also lexical variation (like VerbNet but unlike PropBank), grouping together lexical units of various categories, on the basis of criteria that are not primarily syntactic (unlike VerbNet). Indeed, the FrameNet structure (the set of frames and relations between frames and between roles) has proved to be largely portable to other languages (Boas, 2009) such as Spanish (Subirats-Rüggeberg and Petruck, 2003), Japanese (Ohara et al., 2004), German (Burchardt et al., 2009) or Swedish (Friberg Heppin and Toporowska Gronostaj, 2012) (among others). On top of frames and FEs, frame-semantic corpus annotations have also proved largely portable: Padó (2007) found a high degree of parallelism of the annotations for the English/German and the English/French pairs, in a manually FrameNet-annotated 1000 sentence parallel corpus. Our project aims at producing larger scale manually validated FrameNet resources for French, partially building on previous work that automatically projects the English FrameNet resource to French (Padó, 2007; Mouton et al., 2010).

As far as corpus annotations are concerned, we chose to add semantic annotations on top of manually validated syntactic representations, a strategy that was first adopted within the German SALSA project (Burchardt et al., 2006). We adopted it for two main reasons: first it will provide sentences with both syntactic and semantic gold annotations, which can be useful for corpus studies on the syntax-semantics interface, and which will also be useful for training semantic analyzers. Indeed, it seemed obvious from the very first systems (Gildea and Jurafsky, 2002) that syntactic features are essential clues for predicting semantic frames and FEs. A second motivation is that, whereas the original FrameNet annotations were lexicographic in the beginning (the examples were chosen to maximize the diversity of frame linguistic realizations), we are more concerned with providing a corpus-driven resource that respects "natural" distributions for frames, FEs and their syntactic realizations, usable both for corpus linguistics studies and for semantic analyzer training. This is inspired both by the SALSA project again, and by the more recent full-text annotations performed in the FrameNet project, which have

[1] https://sites.google.com/site/anrasfalda/
[2] https://framenet.icsi.berkeley.edu/

proved to constitute a better training set for semantic analyzers (e.g. (Das et al., 2010)) than the lexicographic annotations.
The two target treebanks for the ASFALDA project are the French Treebank (Abeillé and Barrier, 2004) and the Sequoia treebank (Candito and Seddah, 2012), that contain respectively approximately 18500 sentences[3] and 3000 sentences. The French Treebank is a newspaper corpus, while the Sequoia Treebank was developed, using the same annotation guidelines, with the aim of covering other domains. It contains sentences from a local newspaper, from Europarl, from the French Wikipedia and from the European Medicine Agency.

2. General Strategy

Two main strategies have been proposed in the past for FrameNet developments: a frame-by-frame strategy, that enforces coherence of annotations within a frame, and a lemma-by-lemma strategy that provides annotations reflecting the full ambiguity of a given lemma within the target corpus, a key aspect for the usability of the resource as training data for machine-learning based semantic analyzers. The first strategy is prevalent within FrameNet-related projects, while the second is characteristic of the German FrameNet, and also partly adopted by the Japanese FrameNet.

The frame-by-frame strategy makes it possible to account for the full lexical diversity available to express a frame. Moreover, because a good understanding of the limits of a frame is difficult, this strategy also eases the task of annotators, who can perform better on a frame they know well. Yet, the resulting lexicon is biased: for a given lemma, only senses pertaining to covered frames will appear in the lexicon, and these senses are not necessarily the most frequent senses of that lemma.
The lemma-by-lemma strategy avoids this issue, but requires addressing very diverse lemma senses, often with no existing frame in the English FrameNet database, including rarer senses or cases in which the lemma is part of a larger lexical unit with non fully compositional semantics (Burchardt et al., 2009). While this can be a valid strategy to increase the frame coverage, we draw attention on the substantial difficulty of frame development. For that reason, the SALSA project proposed to cope with senses not covered by any existing English frame by creating sense-specific *proto-frames*, without lexical generalization nor semantic relations to other frames.

Within a three-year project, we cannot reach a coverage comparable to the English FrameNet resource.[4] Yet, in order to maximize usefulness of the rather low-coverage target resource, we have adopted a hybrid strategy:

- In a first phase, we focus on obtaining exhaustive annotation for some *specific notional domains*.
- In a second phase (not started yet), we will perform exhaustive annotation for some lexical units (highly frequent verbs, leaving aside the top n most frequent verbs, to avoid auxiliaries and modals).

With such a hybrid strategy, we hope to obtain a resource that addresses both lexical variation within a frame and semantic relatedness within a notional domain, while also coping with lexical ambiguity, via the lemma-by-lemma annotation strategy. From a practical point of view, the domain-by-domain strategy is quite useful to enforce coherence of frame delimitations, particularly difficult for close frames related to the same broad notion. Preliminary investigations prior to the project revealed that understanding the content of frames and their relations or differences to other frames is quite a difficult task. Working domain by domain mitigates this difficulty, and is perfect in the first phase of the project, since it allows us to train annotators before the even more difficult task of finding or defining frames for every sense of a given verbal lemma (second phase of the project).

3. Delimiting the frames for each notional domain

In the first year of the project, we completed the frame selection (and in some cases adaptation to French) for a set of seven notional domains.

3.1. Selected Domains

- Commercial transaction: originally well studied in the English FrameNet, this domain offers a well circumscribed notion, relatively easy to model. It has the particularity of including converse verbs, for which FrameNet is particularly adapted, when compared to more syntactic-oriented semantic generalizations: the semantic-orientation of frame development allows to use the same semantic role (frame element in FrameNet) for the subject of *to sell* and the indirect object of *to buy*.

- Verbal communication: this notional domain is pervasive in the corpus we plan to annotate, more precisely in the newspaper sentences (the French Treebank, and part of the Sequoia Treebank). It offers representational challenges, for instance in a quotation context, in which the quotation verb is not necessarily conveying *per se* the idea of a communication (Sagot et al., 2010). These authors have studied quotations in news dispatches, showing that a subtantial number of verbs that can appear in parenthetical clauses accompanying direct quotations are not "communication verbs", in the sense that they cannot introduce (be in initial position of) a quotation.

- Judgment/Evaluation: an entity (a person, an object, an event) is evaluated or judged positively or negatively according to a norm. Triggers in this domain have to intrinsically express a judgment or an evaluation, such as *good, bad* or *beautiful*. Lexical items that can be objective in some contexts and evaluative in others are discarded. For instance, the adjective

[3] We use the French Treebank dependency version released for the SPMRL 2013 Shared Task (Seddah et al., 2013).

[4] The 1.5 FrameNet release contains 1019 frames, and about 12000 word-frame associations.

short expresses a negative evaluation in *the phone has a short battery life* whereas it does not express any evaluation in *the girl's skirt is short*.

- Cognitive positions: This notional domain includes predicates where the stance of a cognizer towards a propositional content is expressed. It is mostly concerned with beliefs, with varying degrees of certainty, either stative (*know*, *think*) or inchoative (*realize*). Among the stative beliefs, we retain the FrameNet distinction concerning the beliefs evoked via a lexeme that contains the idea of prediction of a future event (*predict*). We've also included frames referring to closely related concepts such as influence on a cognizer's stance, memory and agreement between cognizers. We consider this domain useful in the perspective of fact extraction. It comprises in particular a large group of factive predicates, which are of primary interest in the perspective of exploiting a FrameNet-Annotated Corpus for Textual Entailment (Burchardt and Pennacchiotti, 2008). In addition, this domain has difficult and interesting interactions with the Judgment/Evaluation domain.

- Spatial relations: Spatial relations are ontologically primitive in most situation descriptions. We currently limit our work to the locative relation frames and the motion frames, leaving aside the body movement frames (rare in our target corpus), and the placement frames (motion causatives). One of the main difficulties in the annotation task will be to differentiate the spatial interpretation of the target expressions from the metaphorical one. We will define spatial entities upon ontological conditions, and assume that located entities (Figure, Theme, Self Mover, etc.) are material entities, phenomenons or events. Landmark entities (Ground, Location, Goal, Path, etc.) are exclusively material entities. The annotation in the Asfalda project may be circumscribed to these spatial cases or include the metaphorical ones, and then use a double annotation pattern, involving a source frame to represent the literal meaning and a target frame to represent the figurative meaning, as performed in the SALSA project (Burchardt et al. 2009).

- Temporal relations: We focus on temporal relations (temporal ordering or inclusion) and duration relations. Aspectual frames (e.g. `Process_start`, `Process_stop`) and temporal frames including causation (e.g. `Change_event_duration`, `Change_event_time`) are left aside.

- Causality: The domain covers both factual causation between events appearing in narratives and evidential or epistemic relations between facts relevant in argumentative texts. We address the most generic causal frames only, some of which subsume a large number of specialized ones.

These domains are not necessarily disjoint, and some of the frames we selected do belong to several domains. We intentionally chose domains that exhibit various degrees of semantic generality, the last three domains being pervasive semantic phenomena. While some of the domains are evoked mainly by verbs or deverbals, spatial, temporal or causal relations can be evoked by prepositions, conjunctions or adverbs too.

3.2. Importance of general domains for discourse semantics

Moreover, the temporal and causal domains have been introduced for their potential to foster progress in challenging issues at the interface between lexical semantics and discourse semantics. Temporal and causal relations are expressed either at the propositional level, within simple clauses, or at the discourse level, as discourse relations between clauses, and texts usually mix both. The lexicon associated to temporal and causal frames in FrameNet is comprehensive and as a result covers both levels: verbs (e.g., *to precede*, *to cause*), nouns and adjectives are most often associated with a semantic contribution to the propositional level, while conjunctions, adverbs and adverbials called discourse connectives (e.g. *then*, *because*, *as a result*) are considered as markers of discourse relations in discourse theories. Texts in which both levels are annotated are uncommon and will prove very valuable for some discourse theories, like SDRT (Asher and Lascarides, 2003), which explicitly posit themselves at the interface between semantics and pragmatics and study the interplay between the two levels. On the one hand, this is to be expected because such theories exploit the semantics of the basic clauses to show how discourse relations emerge, especially when unmarked. On the other hand, this will give new material to make progress in theories that currently assume a clear demarcation between propositional level and discourse level, and are therefore unable to grasp the existing continuum between the two. For instance, causation can be expressed within a clause (*A lightning set off a fire in the building*) or as a connection between two clauses (*A lightning struck the building. As a result, a fire broke out*), but intermediate constructions occur (e.g., *A lightning struck the building, which set off a fire*) and have not been accounted for yet. Corpora annotated with FrameNet will therefore help to finely carve the demarcation line between propositional and discourse levels, or even to question whether such demarcation really makes sense and bring further theoretical developments.

3.3. Frame Selection Methodology

We define the French FrameNet substructure as the subset of frames and frame-to-frame relations pertaining to the notional domains we worked on (it will be later augmented with other frames, when the lemma-by-lemma strategy will be carried on). The basis for the French FrameNet are the frames and frame-to-frame relations as defined in the English FrameNet, release 1.5,[5] which provides us with a substantial amount of work we can build on. While in most cases, the frames in the French FrameNet substructure are exactly those defined in FrameNet release 1.5, we also performed some modifications (see below for the typology of modifications).

[5] https://framenet.icsi.berkeley.edu/fndrupal/framenet_data

Each notional domain was set under the responsability of a team of two or three people (among the authors of this article), who became the "experts" for the FrameNet modelization of that domain. For a given notional domain, we first selected a seed set of frames and/or a seed set of English lexical units obviously related to the domain, and then followed frame-to-frame relations to enlarge the set. For some large domains, we have systematically extracted candidate frames using some relevant FE names (e.g. SPEAKER or COMMUNICATOR for the verbal communication domain, or core CAUSE for the causality domain).

4. Frame remodeling

The FrameNet documentation (Ruppenhofer et al., 2006) details properties that should be stable for lexical units in a given frame, and properties that may vary. So frame delimitations in FrameNet do derive from the English lexical distinctions. While these delimitations are roughly portable across languages (as pointed in the introduction), we sometimes found it necessary to perform "frame remodeling", i.e. to redefine the contours of frames, merge some frames into one, or split frames into several. These changes were either motivated by French/English differences, or simply because the domain-by-domain strategy revealed some very close or redundant frames. We tried to limit the changes in order to maximize the compatibility with the existing huge amount of work (and data) available in the English FrameNet. We detail below the major cases of frame remodeling, and provide quantitative analysis in Table 1. It shows the current number of frames for the various domains, the number of those frames that differ from English, and these numbers broken down by types of differences.

4.1. Frame merging

Indeed, during the frame selection phase, we found that differences between frames were often difficult to judge, even though the "experts" of each domain are linguistically-trained researchers. Each team had both to understand the English FrameNet modelization, and decide for a modelization comptatible with French vocabulary. In order to better understand (and make precise for the future annotators) the exact semantic scope of the various frames, we decided to systematically make explicit the distinctive characteristics of each frame with respect to close frames, within the same domain or in another domain.

In particular, we decided to merge sets of frames for which we had difficulties in defining distinctive characteristics, with the aim of limiting (future) corpus annotation incoherences. We also merged frames that would create artificial polysemy, with lemma senses that would not be considered as different in a monolingual (French) perspective. For instance, in the causality domain, we decided to merge `Cause_to_start` with `Launch_process`. The latter is a frame with just one lexical unit "to launch" and whose only difference with `Cause_to_start` appears to lie in the implicit typing of the cause as an agent, while `Cause_to_start` covers more types of causes, whether agentive or not. While this difference in semantic type for a given frame element can motivate a frame distinction in English, the French translation "lancer" can be used with both agentive or non agentive causes. Such merging decisions are primarily aimed at limiting corpus annotation incoherences.

4.2. Frame-rather-than-FE strategy

For the pervasive notional domains of spatial, temporal and causal relations, we study the three modes offered by FrameNet to account for them and how these should be optimally combined: frames, FEs, and relations between frames.

While relations between frames should clearly be used for stable semantic relations existing between frames (more precisely existing between the sets of lexical units that evoke the two frames), some cases can be annotated both using a frame or a non-essential FE (peripheral or extra-thematic FE in FrameNet terminology). Another source of frame remodeling results from our decision to give priority of annotation via frame over annotation as filler of FE. Let us take the example of a prepositional locative modifier like in *He died in Europe*. The FrameNet project provides two competing ways of annotating the locative information: either with the *in* preposition evoking its own locative frame (the `Locative_relation` frame), or with the whole PP as filling a non-essential frame element for the frame evoked by *died* (the Place extra-thematic FE of associated with many frames, among which the `Death` frame). While the FrameNet guidelines suggest that the latter annotation type (using an extra-thematic FE) can be interpreted as a shorthand for the former (using a frame)[6], we choose for the French FrameNet annotations to systematically trigger a frame annotation whenever possible, and thus not to annotate extra-thematic FEs, nor FEs that are not encoded as extra-thematic in FrameNet, but that are systematically evoked using a preposition, adverb or conjunction whose meaning can trigger a frame by itself (this is the case for the `Area` FE of the motion frames, see section 4.).

This strategy has sometimes lead to remodel FEs. For instance, in order to uniformely treat localization PPs as triggering the Locative_relation frame, we had to remove the Area FE in the motion frames. Indeed, `Area` is a core FE in motion frames and corresponds to the place where the motion occurs. In a sentence such as *John is walking in the park*, *in the park* refers to the location of John's walk, and it would be annotated as `Area` in the English FrameNet. In Asfalda, we do not distinguish between static localizations of motions and static localizations of any other eventuality, so for the French equivalent *Jean marche dans le parc*, the location (*le parc*) will be annotated as the `Ground` FE of the `Locative_relation` frame triggered by the preposition *dans ('in')*, and the Area role is not taken into account in the `Motion` Frame. For the verbs that only convey static localization, like in *La clé est/se trouve/se situe derrière la*

[1] Since some frames belong to several domains, the total number is less than the sum of numbers of frames for each domain.

[6] (Ruppenhofer et al., 2006), p. 97. Note that this redundancy can be explained by the lexicographic strategy adopted within the FrameNet project: when annotating a chosen example for the `Death` frame, annotating the Place information is faster than instantiating an additional `Locative_relation` frame.

Notional domain	Nb. of frames	Total nb. modified frames	Merges	Splits	Modifications of roles only	Other modifications
Commercial transaction	11	0	0	0	0	0
Verbal communication	34	3	2	1	0	1
Judgement/Evaluation	16	0	0	0	0	0
Cognitive positions	19	4	2	1	0	1
Temporal relations	4	2	1	0	0	1
Spatial relations	20	14	2	0	12	0
Causality	8	5	1	0	4	0
Total	106[1]	28	8	2	16	3

Table 1: Current number of (lexicalized) frames in French FrameNet substructure, for each notional domain, with number of modified frames with respect to the English FrameNet frames (in total, and divided into difference types).

porte (John is/is located/is situated behind the door), we consider that the preposition triggers the localization frame, while the verb doesn't trigger any frame. The homogeneity of treatment of all expressions denoting static localization is thus preserved.

4.3. Frame splitting

In some cases, we decided to split frames into two French-specific subframes. The Suasion frame encompasses a situation in which an agent deliberately communicates a message to a cognizer, with the aim of either influencing the cognizer's stance on the veracity of a content, or on his readiness to perform an action. Looking at this frame's triggers, we found that some of their French translations could only refer to cases in which the agent's influence pertained to the cognizer's willingness to act (*décider, dissuader*). This and the fact that a splitting operation is easily reversible and can only provide additional information made us decide on splitting Suasion into two subframes whose difference would be their Content FE's semantic type (action vs. content).

4.4. Miscellanous

A special case of frame modelization modification pertains to the account of temporal modifiers. In case of prepositional temporal modifier, for instance in (1), the preposition evokes the `Time_vector` or `Temporal_collocation` frame, and the object of the preposition (*december 3rd*) fills the `Landmark` FE. Additionnally, *december* triggers the `Calendar_units` frame. Now, in the case of a direct temporal modifier, as in (2), while the units *yesterday, day, year* do evoke the `Calendar_units` frame, the temporal collocation relation itself is not currently captured, due to the absence of preposition.

(1) *Paul sold his car on december 3rd.*

(2) *Paul sold his car yesterday / that year / the very day of his wedding...*

The French counterparts of these two examples have the same structure. One possibility to account for the temporal relation in (2) could be to consider that *yesterday, day, year* can trigger the temporal collocation frame, while incorporating the Landmark FE. This option is not optimal since the full expression for the Landmark can comprise modifiers (e.g. *the very day of his wedding*). When switching to French, we get the same problems for the French counterparts of (2), and an additional problem for (1'), the French counterpart of (1), because dates can be direct modifiers. Having the month *décembre* trigger the temporal collocation frame and incorporate the Landmark would not account for the full Landmark date.

(1') *Paul a vendu sa voiture le 3 décembre.*
Paul has sold his car the 3 december.
(Paul sold his car **on** *december 3rd.)*

This leads us either to assume an empty trigger of the temporal collocation frame, or more appropriately to consider that the temporal relation is conveyed by the modifier construction (in the same vein as the constructicon (Fillmore et al., 2012)). Yet for practical reasons, we cannot adopt either of these solutions. In case of direct temporal modifier, we will consider the whole modifier as both triggering the temporal collocation and the Landmark. incorporation.

5. Lexicon development

One main objective of the Asfalda project is to provide a free corpus-driven syntactico-semantic lexicon, that will be extracted from the semantic annotations on treebanks (and will thus be biased by our two target corpus, but informed with numbers of occurrences of various phenomena). Yet, because the semantic modelisation via frames and semantic annotation is a difficult task, we've worked in a first phase on validating a French FrameNet lexicon, that associates lemmas to frames. This lexicon will then be used to guide the semantic annotation on treebanks, domain by domain, and in return, a corpus-driven lexicon will be re-extracted from the annotations.

5.1. Lexicon validation tool

The lexicon development prior to annotations was performed via the Asfalinks tool, specifically developed for that purpose. The Asfalinks tool is a Qt C++ cross-platform application using the Subversion tool library to handle its multi-users features. It allows to edit, annotate and adjudicate the frame evoking elements of frames. Figure 1 shows the adjudication mode of the lexicon. The lemma *abriter.v* has been validated by all annotators as frame-evoking element for the *Containing* frame, while *cacher.v* has been

refused. The lexical unit *comporter.v* has conflicting status, which must be resolved by the adjudicator.

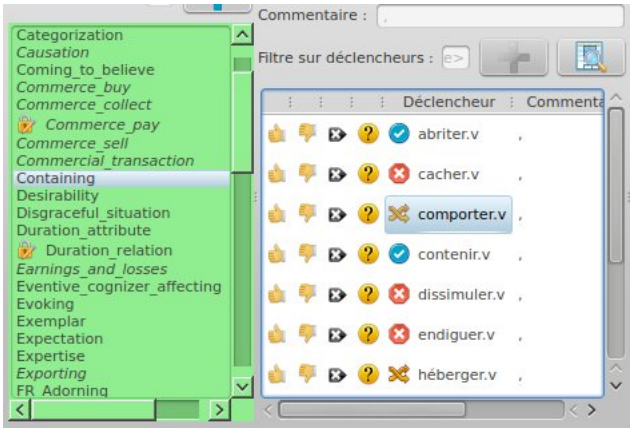

Figure 1: Snapshot of lexicon adjudication, via the Asfalinks GUI.

Validators and adjudicators also had access to a KWIC concordancer, to easily spot and study occurrences of a lemma in our target treebanks.

5.2. Lexicon validation methodology

Once each team had selected the frames for a given domain, made explicit their distinctive characteristics, and performed an initial frame remodeling, we built the corresponding French FrameNet lexicon, using three steps we detail below, which have in return lead to modifications in the frames, and refining in the frame distinctive characteristics.

5.2.1. Initial lexicon development

For each frame of the French substructure, two *experts* of the domain (the people in charge of delimiting the frames for the domain) independently validated lexical units, using the Asfalinks tool. The lexical units proposed for a frame originated from the merge of two automatically built French FrameNet lexicon, by transfer from the English FrameNet lexicon: one using word alignments from parallel corpora (Padó and Lapata, 2005) and one using bilingual dictionaries and filtering heuristics (Mouton et al., 2010).[7] For each proposed entry, the validators were asked to choose between valid and invalid, and a third 'unsure' status, to use in case of hesitation, in order to force the entry to be examined at adjudication time. Validators were also asked to enter missing entries if necessary (or to build from scratch in case of remodeled frame). They used synonym dictionaries (in particular the DES[8]) to obtain additional candidates. They also relied on additional domain-specific lexicons: Lexconn (Roze et al., 2012) for the Causality domain, and for the Judgment/Evaluation domain, a subjective French lexicon developed within the CASOAR project that aimed at measuring the impact of discourse structure on opinion analysis (Benamara et al., 2011; Chardon et al., 2013). Time related frames benefited from work on TimeML compliant analyzers for French (Parent et al., 2008; Bittar et al., 2011).

5.2.2. Adjudication

For each domain, the same two experts performed adjudication together, except for cross-domain frames, which were adjudicated by experts for each relevant domain. This task has often lead to enrich and clarify the distinctive characteristics for some pairs of frames, and also in some cases to modify the frame lexical scopes.

5.2.3. Polysemy checking

Because the two first steps revealed to be quite a difficult task, we decided to further enforce coherence of the resulting lexicon by checking its *internal* polysemy, namely all the cases of lemmas associated to several frames in the French FrameNet substructure (namely 282 lemmas). The instruction for this task was to provide an effective disambiguating test for any polysemic LU or more generally for any pair of frames sharing numerous lemmas. The result of this phase both provides disambiguating tests to be used by annotators during the corpus annotation phase (section 6.), and in some cases revealed incoherences or redundancies between frames belonging to different domains/ It thus lead to clarify frame distinctions and modify the lexicon accordingly (we added 49 and removed 46 lexeme/frame associations).

We end up with a French FrameNet lexicon covering 106 lexicalized frames, and totalizing 2244 lexeme/frame pairs, corresponding to 1936 distinct lemmas. When considering only lemmas appearing at least once in our target treebanks, we obtain 1638 lexeme/frame pairs (corresponding to 1359 distinct lemmas, among which 797 are verbs, 574 nouns, 270 adjectives, 41 conjunctions and 119 prepositions).
Concerning the evaluation of that resource, computing precise inter-annotator agreement does not make much sense. Indeed, as we already suggested, the initial definition of frames for each domain was iteratively modified during the lexicon development, in particular at adjudication time and later on during polysemy checking. This is not very surprising given the highly lexical nature of frames. We believe that working domain-by-domain, and with domain experts, actually helped to increase the resource quality.

6. Preliminary Work for Corpus Annotation

The next phase of the project is frame and FE annotation on dependency trees of the French Treebank and the Sequoia Treebank. For each lemma of the lexicon, occurrences of lemmas (up to a certain number of occurrences per lemma) will be pre-annotated with all the frames associated to the lemma. External annotators will then be asked to choose whether the meaning in context corresponds to one of the proposed frames, and if so to annotate the FEs. An example is shown in section 2, for sentence (3), for which two frames are proposed for the verb *encourager (encourage)*. Annotators will be provided with the disambiguation guides written during the polysemy checking phase. They will not have the possibility to directly add new frames for

[7] In case of a remodeled frame, the entries of close English frames were used as starting point to validation.

[8] http://www.crisco.unicaen.fr/des/

a given lemma, as this will be submitted to validation by the domain experts (and these cases should be limited given the preliminary work on the lexicon that was done by the domain experts).

(3) *L' évolution spontanée des taux les y encourage*
The evolution spontanée of-the rates them to-it encourages.
(The spontaneous evolution of the rates encourages them to do so)

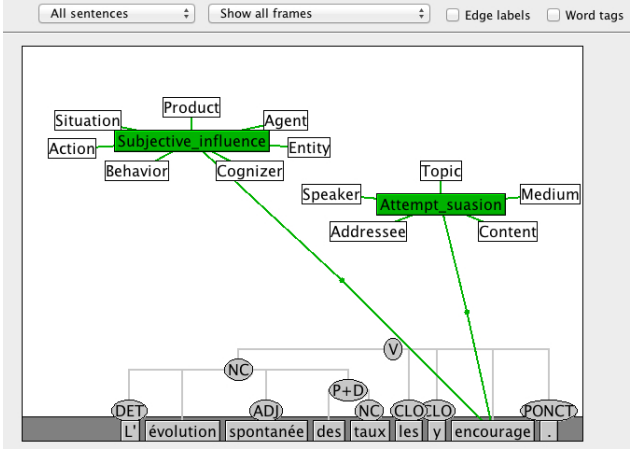

Figure 2: Snapshot of the planned annotation on a dependency tree, via the Salto tool, for sentence (3)

7. Conclusion

We presented the current status of the French FrameNet development. Focusing on six notional domains, we delimited a set of 106 frames, adapted from the English FrameNet frames, and defined the corresponding French lexicon. We could verify that the domain-by-domain methodology we adopted helped to enforce the consistency of frame/lexemes associations. The next phases of the project concern frame and FEs annotation on treebanks, using automatic pre-annotation.

Acknowledgements

This work was funded by the French National Research Agency (ASFALDA project ANR-12-CORD-023), and supported by the French Investissements d'Avenir - Labex EFL program (ANR-10-LABX-0083).

8. References

Anne Abeillé and Nicolas Barrier. 2004. Enriching a french treebank. In *Proc. of LREC'04*, Lisbon, Portugal.

Nicholas Asher and Alex Lascarides. 2003. *Logics of Conversation*. Cambridge University Press, Cambridge.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *COLING-ACL '98: Proceedings of the Conference*, pages 86–90, Montreal, Canada.

Farah Benamara, Baptiste Chardon, Yvette Yannick Mathieu, and Vladimir Popescu. 2011. Towards context-based subjectivity analysis. In *IJCNLP*, pages 1180–1188.

André Bittar, Pascal Amsili, Pascal Denis, and Laurence Danlos. 2011. French TimeBank: An ISO-TimeML Annotated Reference Corpus. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (short papers)*, pages 130–134. Association for Computational Linguistics.

Hans Christian Boas, editor. 2009. *Multilingual FrameNets in computational lexicography : methods and applications*. Trends in linguistics. Mouton de Gruyter, Berlin, New York.

Aljoscha Burchardt and Marco Pennacchiotti. 2008. FATE: a framenet-annotated corpus for textual entailment. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2008/.

Aljoscha Burchardt, Katrin Erk, Anette Frank, Andrea Kowalski, Sebastian Padó, and Manfred Pinkal. 2006. The salsa corpus: a german corpus resource for lexical semantics. In *Proceedings of LREC 2006*, Genoa, Italy.

A. Burchardt, K. Erk, A. Frank, A. Kowalski, S. Padó, and M. Pinkal. 2009. Framenet for the semantic analysis of german: Annotation, representation and automation. In Hans C. Boas, editor, *Multilingual FrameNets in Computational Lexicography: Methods and Applications*, volume 200 of *Trends in Linguistics*, pages 209–244. Mouton de Gruyter.

Marie Candito and Djamé Seddah. 2012. Le corpus sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. In *Proc. of TALN 2012 (in French)*, Grenoble, France, June.

Baptiste Chardon, Farah Benamara, Yvette Yannick Mathieu, Vladimir Popescu, and Nicholas Asher. 2013. Measuring the effect of discourse structure on sentiment analysis. In *CICLing (2)*, pages 25–37.

Dipanjan Das, Nathan Schneider, Desai Chen, and Noah A. Smith. 2010. Probabilistic frame-semantic parsing. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, HLT '10, pages 948–956, Stroudsburg, PA, USA.

Charles Fillmore, Russel Lee-Goldman, and Russel Rhomieux. 2012. The framenet constructicon. In Boas Hans and Sag Ivan, editors, *Sign-Based Construction Grammar*, pages 309–372. Stanford: CSLI.

Karin Friberg Heppin and Maria Toporowska Gronostaj. 2012. The rocky road towards a swedish framenet - creating swefn. In *Proceedings of the Eighth conference on International Language Resources and Evaluation (LREC-2012); Istanbul, Turkey*, pages 256–261.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288, September.

Claire Mouton, Gaël de Chalendar, and Benoît Richert. 2010. Framenet translation using bilingual dictionaries with evaluation on the English-French pair. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta,

Malta, may. European Language Resources Association (ELRA).

Kyoko Ohara, Seiko Fujii, Shun Ishizaki, Toshio Ohori, Hiroaki Saito, and Ryoko Suzuki. 2004. The Japanese FrameNet project; an introduction. In Charles J. Fillmore, Manfred Pinkal, Collin F. Baker, and Katrin Erk, editors, *Proceedings of the Workshop on Building Lexical Resources from Semantically Annotated Corpora*, pages 9–12, Lisbon. LREC 2004, LREC 2004.

Sebastian Padó and Mirella Lapata. 2005. Cross-lingual bootstrapping for semantic lexicons: The case of framenet. In *Proceedings of the 20th National Conference on Artificial Intelligence*, pages 1087–1092, Pittsburgh.

Sebastian Padó. 2007. *Cross-Lingual Annotation Projection Models for Role-Semantic Information*. Ph.D. thesis, Saarland University. MP.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106, March.

Gabriel Parent, Michel Gagnon, and Philippe Muller. 2008. Annotation d'expressions temporelles et d'événements en francais. In *TALN 2008*. ATALA.

Charlotte Roze, Laurence Danlos, and Philippe Muller. 2012. LEXCONN: a French lexicon of discourse connectives. *Discours*, 10.

Josef Ruppenhofer, Michael Ellsworth, Miriam R.L. Petruck, Christopher R. Johnson, and Jan Scheffczyk. 2006. *FrameNet II: Extended Theory and Practice*. International Computer Science Institute, Berkeley, California. Distributed with the FrameNet data.

Benoiît Sagot, Laurence Danlos, and Rosa Stern. 2010. A lexicon of french quotation verbs for automatic quotation extraction. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

Karin Kipper Schuler. 2005. *Verbnet: a broad-coverage, comprehensive verb lexicon*. Ph.D. thesis, Philadelphia, PA, USA. AAI3179808.

Djamé Seddah, Reut Tsarfaty, Sandra K´ubler, Marie Candito, Jinho Choi, Richárd Farkas, Jennifer Foster, Iakes Goenaga, Koldo Gojenola, Yoav Goldberg, Spence Green, Nizar Habash, Marco Kuhlmann, Wolfgang Maier, Joakim Nivre, Adam Przepiorkowski, Ryan Roth, Wolfgang Seeker, Yannick Versley, Veronika Vincze, Marcin Woliński, Alina Wróblewska, and Eric Villemonte de la Clérgerie. 2013. Overview of the spmrl 2013 shared task: A cross-framework evaluation of parsing morphologically rich languages. In *Proceedings of the 4th Workshop on Statistical Parsing of Morphologically Rich Languages: Shared Task*, Seattle, WA.

Carlos Subirats-Rüggeberg and Miriam R.L. Petruck. 2003. Surprise: Spanish FrameNet! In Eva Hajičová and Anna Kotéšovcová and Jiří Mirovský, editor, *Proceedings of the Workshop on Frame Semantics, XVII International Congress of Linguists (CIL)*, Prague. Matfyzpress, Matfyzpress.