Corpus annotation within the French FrameNet: a domain-by-domain methodology

Marianne Djemaa, Marie Candito, Philippe Muller, Laure Vieu

Corpus annotation within the French FrameNet: a domain-by-domain methodology

Marianne Djemaa*, Marie Candito*, Philippe Muller$^{\diamond}$, Laure Vieu$^{\triangle}$
$\star$ Alpage (Univ. Paris Diderot / INRIA), $\diamond$ IRIT, Toulouse University , $\triangle$ IRIT, CNRS,
asfalda@inria.fr

Abstract

This paper reports on the development of a French FrameNet, within the ASFALDA project. While the first phase of the project focused on the development of a French set of frames and corresponding lexicon (Candito et al., 2014), this paper concentrates on the subsequent corpus annotation phase, which focused on four notional domains (commercial transactions, cognitive stances, causality and verbal communication). Given full coverage is not reachable for a relatively "new" FrameNet project, we advocate that focusing on specific notional domains allowed us to obtain full lexical coverage for the frames of these domains, while partially reflecting word sense ambiguities. Furthermore, as frames and roles were annotated on two French Treebanks (the French Treebank (Abeillé and Barrier, 2004) and the Sequoia Treebank (Candito and Seddah, 2012), we were able to extract a syntactico-semantic lexicon from the annotated frames. In the resource's current status, there are 98 frames, 662 frame-evoking words, 872 senses, and about 13000 annotated frames, with their semantic roles assigned to portions of text. The French FrameNet is freely available at `alpage.inria.fr/asfalda`.

Keywords: FrameNet, French, semantic roles, semantic frames, semantically-annotated corpus

1. Introduction

The ASFALDA project[1] aims to build semantic resources and a corresponding semantic analyzer for French, to capture generalizations both 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), which provides a structured set of prototypical situations, called **frames**, along with a semantic characterization of the participants of these situations (called *frame elements*, but we'll use **roles** for short). The corresponding English lexicon associates frames with the words that can evoke them (called *frame-evoking elements*, **FEEs** for short). 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 over not only syntactic variation (e.g. diathesis alternation) but also lexical variation (like VerbNet but unlike PropBank), and groups together lexical units of various categories, on the basis of criteria that are not primarily syntactic (unlike VerbNet).

The resources built within ASFALDA consist of a set of frames, 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 project also aims to investigate new models for frame-based semantic analysis. In a first phase of the project (Candito et al., 2014), the work focused on a set of notional domains, and frames pertaining to these domains were selected from the frame set of Berkeley FrameNet 1.5. These frames were adapted to French, and the corresponding French lexicon was built. The current paper focuses on the subsequent corpus annotation phase.

In section 2., we describe the original FrameNet developed for English. In section 3., we review the pros and cons of various FrameNet building methodologies before presenting ours, and then describe our target corpora and our annotation workflow. We next present the resulting resource in section 4., evaluating its quality using inter-annotator agreement and providing various statistics. We then review specific phenomena we had to address in section 5., and conclude in section 6..

2. FrameNet

FrameNet (Baker et al., 1998), developed at UC Berkeley's ICSI, is a Fillmore (1982)'s Frame Semantics inspired resource. It is made of: (a) a network of **frames**: prototypical scenes, complete with semantic characterizations of the scene's participants, props or parts called frame elements (as mentioned in the introduction, we will say **roles** for short); (b) a lexicon of lexemes that may evoke those frames; and (c) a corpus of annotated English sentences. These sentences carry frame annotations on the words that evoke them, as well as role annotations on the linguistic material that realizes those roles.

As we can see in figure 1 (which shows excerpts from the EXPORTING frame), a frame consists of several parts, the first of which is a natural language definition of the prototypical situation described by the frame. Each frame also includes a set of roles − those may be assigned specific semantic types, and relations (such as exclusion) may be defined between several roles. Frames are linked via frame relations, such as inheritance or is-causative-of. A key aspect of FrameNet is that roles are specific to each frame, as an answer to the well-known difficulty of fitting the semantic arguments of predicates into a small set of semantic roles (Fillmore, 2007). Yet, more coarse-grained roles can be obtained through frame-to-frame relations: identity relations are defined between the roles of two frames linked via a frame-to-frame relation. This makes it possible to generalize over frame-specific roles to obtain a granularity closer

[1] alpage.inria.fr/asfalda

EXPORTING		
Def:	An *Exporter* moves *Goods* across a border from an *Exporting_area* to an *Importing_area.*	
Roles:	*Exporter*	The conscious entity, generally a person, that moves the *Goods* across a border out of the *Exporting_area*
	Exporting_area	The place where the *Goods* are initially, before the the *Exporter* moves them.
	Goods	The items of value whose location is changing.
	Importing_area	The place that the *Goods* end up as a result of motion.
FEEs:	export.n, export.v, exportation.n	

Figure 1: EXPORTING frame

to that of traditional sets of semantic roles.
Roles are grouped by status, based on whether they are essential to the meaning of the frame. In particular, roles are called "core" if they "instantiate a conceptually necessary component of a frame, while making the frame unique" (Ruppenhofer et al., 2006)[2].
In every frame is also listed the set of lexical units that may evoke it: the **frame-evoking elements** (FEEs), made of a lemma and a part-of-speech. Each frame and FEE pair usually comes with a choice of annotated sentences from the British National Corpus, picked so that they exhaustively examplify the range of possible role syntactic realizations of the roles of a frame. Each such examplar contains one **annotation set** exactly, namely one frame annotated on the FEE along with the various roles annotated on the corresponding role fillers. In a later stage of the project, full-text annotations were added, which consist of a set of running texts in which every occurrence of potentially frame-bearing words has been annotated.
Compared to other semantic roles resources, FrameNet has a more semantic orientation[3]. Yet FrameNet annotation is still lexicon-based: two sentences referring to a same situation might bear different frames depending on the words used.
Initiated in 1998, the resource currently has 13 474 FEEs that may evoke a total of 1 216 frames. There are 174 038 annotation sets in the lexicographic corpus, and 28 046 in the full-text corpus[4]. The French FrameNet was designed using the slightly smaller 1.5 release.
Probably because of its semantic orientation, FrameNet has been shown to be quite portable to other languages (Boas, 2009), and FrameNet-like resources have already been developed for, among others, Spanish (Subirats-Rüggeberg and Petruck, 2003), Japanese (Ohara et al., 2004), German (Burchardt et al., 2006b) and Swedish (Borin et al., 2009).

[2]Although primarily stated in semantic terms, coreness is closely related to the formal distinction of argument versus adjunct of the underlying FEEs. Note though that some subcategorized complements may not be conceptually necessary, such as the addressee of the Statement frame.

[3]For instance unlike VerbNet (Schuler, 2005), which is based on Levin's verb classes (Levin, 1993), FrameNet's criteria for frame delimitation and role typing are mostly semantic.

[4]https://framenet.icsi.berkeley.edu/fndrupal/current_status, consulted in march 2016

3. Methodology

3.1. Pros and cons of existing frame semantics annotation strategies

The original Berkeley FrameNet project and subsequent FrameNet projects for various languages have adopted various methodologies, which have a major impact on the resulting resources: because full lexicon coverage is unreachable, the path taken to carry out the annotations shapes the result. We therefore describe the main traits of previous methodologies, before presenting ours.
Three main strategies have been used in the past:

- The lexicographic frame-by-frame strategy is prevalent within FrameNet-related projects. A frame is defined along with its FEEs, and examplar sentences containing (disambiguated) occurrences of these FEEs are chosen for annotation, with the objective of maximizing the range of annotated syntactic valences for a given semantic frame. Each examplar sentence contains one annotation set only.

- The corpus-driven lemma-by-lemma strategy is characteristic of the German FrameNet (developed in the SALSA project, (Burchardt et al., 2009)), and was also partly adopted by the Japanese FrameNet (Ohara et al., 2004). A set of lemmas is chosen[5] and all their occurrences in the target corpus are annotated.

- The full-text strategy, which was later adopted within the Berkeley FrameNet project, consists in annotating any content-word occurrence in running text. Though it presupposes the existence of a core set of frames, it necessarily entails defining new frames as uncovered senses are encountered.

This last strategy obviously achieves perfect coverage, though restricted to the target corpus, in terms of the lexical diversity for a given frame and of the sense ambiguity for a given lemma. Yet it is extremely difficult to achieve: it is likely that the full-text annotations could only be completed thanks to the experience acquired through the lexicographic phase of the Berkeley FrameNet project. Preliminary trials led us to the conclusion that full-text frame semantic annotation could not be accomplished within our project.
The other two strategies have different flaws and qualities. The frame-by-frame annotation enforces full lexical coverage for frames, and fetching lexicographic examples enforces full coverage of the possible syntactic variation within a frame. Yet because frame coverage is necessarily insufficient, the resulting lexicon is biased: for a given lemma, only senses pertaining to covered frames will appear in the lexicon, even though these senses may not be the most frequent senses of that lemma. On the contrary, with the lemma-by-lemma strategy, lemmas are either fully

[5]In the SALSA project, the target lemmas were primarily 500 verbs of all frequency bands, plus some deverbal nouns.

accounted for or entirely missing, and thus the lexical diversity of a frame is not fully accounted for.

As far as ease of annotation is concerned, since achieving a good understanding of the limits of a frame is quite a difficult task, the frame-by-frame strategy eases the task of annotators, who can perform better on a frame they know well. The lemma-by-lemma strategy requires addressing very diverse lemma senses, often without an existing frame in the Berkeley 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 frame coverage, preliminary investigations led us to conclude it was extremely difficult for an annotator to master frames from very various notional domains, and *a fortiori* to be able to create new frames[6].

On the other hand, we have experienced that using the frame-by-frame strategy, i.e. working in isolation on a frame, can result in missing some similarities with other frames, and thus artificially increasing polysemy: choosing the right frame for a given lemma's occurrence can become quite difficult due to blurred frontiers between frames.

Finally, a crucial characteristic of the different strategies concerns the use of the resulting annotations as training data for semantic parsers: even though they are much more numerous, Berkeley FrameNet's lexicographic annotations have proved much less useful than their full-text ones (Das et al., 2010), which have the crucial trait of preserving the natural sense and role-realization probabilistic distributions.

3.2. Corpus-driven domain-by-domain strategy

From these observations, we chose to use a "corpus-driven domain-by-domain strategy", i.e. to focus on a small set of notional domains and fully annotate the frames pertaining to these domains. Within a given domain, we aimed for full coverage in terms of the FEEs that can evoke the frames, and full coverage in terms of the occurrences of such FEEs within a target treebank. In a first phase of the project (Candito et al., 2014):

- We chose a small set of notional domains, and selected the English frames pertaining to these domains (within the 1.5 FrameNet release).
- We worked in parallel to (i) define the lexicon for these frames, starting from automatically projected French FrameNet lexicons (Mouton et al., 2010; Padó, 2007); (ii) adapt the frames to French and (iii) cope with frame overlap brought to light thanks to the domain-by-domain approach. This has sometimes led us to merge several English frames into one. Some frames were split and totally new frames were created, to complete the frame modelization of a notional domain.

[6]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.

In the current paper, we describe the subsequent corpus annotation phase. We used as target corpus two syntactically-annotated corpora (see below section 3.3.), and basically aimed to annotate all occurrences of lemmas that *potentially* evoke a frame of one of 4 notional domains (cf. section 4.1.). Once disambiguated, the FEE occurrence is either associated with one relevant frame, or with a special Other_sense frame, to indicate the meaning is outside of the targetted notional domains. More precisely, we used an upper bound of 100 occurrences[7] of a given lemma. This methodology entails that an occurrence of a lemma appearing in our lexicon is either (i) annotated with a frame from one of the 4 notional domains, (ii) annotated with the dummy Other_sense frame or (iii) not associated to any frame at all (meaning it was not proposed for annotation at all because it was beyond the first 100 occurrences).

In the resulting resource, such a strategy entails that:

- a given notional domain is completely annotated;
- a frame is associated to all its potential FEEs, and their occurrences (if any) in the corpus are frame-annotated. We thus retain the crucial property of the FrameNet project of capturing the lexical diversity within a frame, including diversity of parts-of-speech;
- sentences have both full syntactic gold annotation and partial semantic annotations, which can be useful for corpus studies on the syntax-semantics interface;
- the resulting annotations do reflect naturally occurring sense and role distributions, though only within the four target notional domains, and modulo the 100-occurrence upper bound. The resource can thus be used as training data for partial word sense disambiguation, and for semantic role labeling, for all the occurrences associated to a frame (whether actual or dummy).

3.3. Target corpora

The corpus we used is the concatenation of two syntactically annotated treebanks: the Sequoia treebank (Candito and Seddah, 2012) and the French Treebank (Abeillé and Barrier, 2004), totalizing 21634 sentences and 624187 tokens. The French Treebank consists of sentences from *Le Monde* national newspaper. The version we used, from the SPMRL 2013 shared task (Seddah et al., 2013), contains 18535 sentences. The Sequoia treebank was developed using the same annotation scheme. It is much smaller (3099 sentences), and contains sentences from 4 different sources[8] : Europarl, a regional newspaper L'Est Républicain, public assessment reports for medicines from the European Medicines Agency, and the French

[7]Actually the first 100 occurrences in the Sequoia plus French Treebank corpus, in that order.

[8]The Sequoia treebank was originally developed to provide "out-of-domain" sentences for parsing experiments. It is important to note that we use in this paper the term "domain" in the sense of "notional domain", different from the "human activity domain" sense that can be used to qualify corpora. The two corpora we used have different origins, but both contain occurrences of FEEs evoking one of our 4 notional domains.

Wikipedia.

While both corpora are available with both phrase-structure trees and dependency trees, we chose to add the semantic annotations on top of the syntactic dependency trees. These syntactic annotations were essential to the goal of obtaining a syntactico-semantic resource, and also as a means to speed up annotations, mainly in the (good) cases in which a role filler coincides with a syntactic subtree.

3.4. Annotation workflow

As stressed in section 3.2., we chose to focus on a set of notional domains in order to define the French frames and their corresponding FEEs. Yet, the ambiguity of any given lemma occurrence forces us to adopt a lemma-by-lemma corpus annotation workflow.
Six annotators worked in this phase, plus 4 domain experts (the authors of this paper). The experts first performed some trial annotations, in order to set up a first version of the annotation guide. Annotation then started for lemmas associated with one frame only, then lemmas associated with several frames but within the same notional domain at first before switching to cross-domain ambiguous lemmas[9]. The annotation guide was continuously expanded based on the annotators' feedback and questions.
More precisely, for each given lemma to annotate, the following steps were applied:

- **Pre-annotation with ambiguous frames**: the first 100 occurrences of the lemma are automatically pre-annotated with its corresponding candidate frames according to the lexicon.
- **Double annotation on top of syntactic dependency trees**: two annotators then have to independently annotate these occurrences, using the Salto graphical tool (Burchardt et al., 2006a): for a given pre-annotated occurrence, each annotator decides whether the occurrence evokes one of the proposed frames. If so, they discard all other frames, and annotate the relevant roles. Only the core roles were to be annotated, as well as non-core roles realized as a subcategorized complement (cf. section 2.). As noted above, the syntactic structure helped speed up the annotation of role fillers when they coincided with a subtree. Note though that annotators could in any case choose the exact tokens composing a role filler, independently of the provided syntactic analysis. This is crucial for cases of syntax/semantics discrepancies and for cases of errors in syntactic annotations.

  For lemmas they found difficult to disambiguate, the annotators could ask "domain experts" to provide a specific disambiguation guide. This could result in modifying the lexicon (i.e. the frames associated to the lemma in the lexicon). When annotating the frames for a predicative noun, annotators were also asked to add special frames for support verbs if needed (these were not pre-annotated). Annotators had the possibility to leave several frames if they were unsure.
- **Adjudication**: the two annotated versions are adjudicated, either by an expert, or by the two annotators together.[10]

[9] Note that in any case, additional ambiguity could arise from senses not belonging to the target notional domains.

4. Resulting annotated resource

4.1. Annotated notional domains

Though the first phase of the project focused on 7 notional domains, we were able to annotate only 4 of these:

- **Commercial transactions** : originally well studied in the English FrameNet, this domain has the particularity of including converse verbs, for which FrameNet is particularly adapted;
- **Cognitive positions**: This notional domain includes predicates in which the stance of a cognizer towards a propositional content is expressed. It is mostly concerned with beliefs, with varying degrees of certainty, and either stative (*know*, *think*) or inchoative (*realize*).
- **Causality**: The domain covers both factual causation between events appearing in narratives and evidential or epistemic relations between facts relevant in argumentative texts.[11]
- **Verbal communication**: this notional domain is pervasive in the journalistic parts of our target corpora. In the current release, it is not fully annotated though: we started with the FEEs ambiguous with the other 3 domains.

A frame generally pertains to one domain only, but not always. For instance the FR_Attributing_cause frame (when someone attributes an effect to a cause) relates both to the causality and the cognitive stances domain.

4.2. Evaluation of the annotation task : Inter-annotator agreement

As noted by (Burchardt et al., 2009), chance corrected inter-annotator agreement metrics such as the kappa score are applicable to classification tasks in which both the items to classify and the classes are fixed. This is not the case in our setting: for the frame assignment task, items to annotate were almost fixed (except for support verbs), but the frames could vary. For the role assignment task, items to add a role to were not fixed, and the set of roles depended on the frame assigned by the annotator. We thus evaluated the annotation task simply using an inter-annotator Fscore, both for frame assigment and role assignment.[12] For the

[10] The first adjudications were performed by the experts plus the annotators, in order to train them. Then, adjudications were performed by annotators if the inter-annotator agreement was high enough (see section 4.2.), by experts otherwise.

[11] We address the most generic causal frames only, some of which subsume a large number of specialized ones.

[12] The Fscore is the harmonic mean of precision and recall when considering one annotation as the reference and the other as the prediction. Interverting both annotations swaps precision and recall values, resulting in the same Fscore.

	Nb of FEE occ.	% of N	% of V	Inter-annotator Fscore		
				Frame	Exact Role	Partial Role
	17667	36	50	85.9	77.2	81.9
Break-down by notional domain[13]						
Commercial	3307	60	40	92.0	73.4	80.4
Causality	7691	30	48	79.2	74.2	80.4
Cog. Stances	7886	28	62	90.6	81.1	86.0
Communic.	2221	23	76	89.6	82.3	87.5
Break-down by POS of the FEE						
V	8834	-	-	87.6	82.8	87.1
N	6234	-	-	86.8	68.3	72.5
other	2509	-	-	77.7	74.6	82.1

Table 1: Inter-annotator agreement for all occurrences of FEE that were independently annotated by two annotators, in total and broken down by notional domain, and by part-of-speech of the FEE. First col.: number of FEE occurrences proposed for double annotation. Next two col.: percentage of nouns and of verbs within the FEE occurrences to annotate. Last three col.: Inter-annotator agreement F_1 scores for the frame assignment, exact role assignment, and partial role assignment tasks (for matching frames).

latter, we report both exact and partial match Fscores. For exact match Fscore, the number of common role fillers are the pairs of role fillers with both same role and exact same set of tokens. For partial match Fscore, we also give partial credit if two role fillers sharing some tokens have been assigned the same role: in order to calculate the amount of correct answers, for each role filler in annotation 1, we look for a role filler in annotation 2 with both same role and highest intersection over union ratio, and add this highest ratio to the total of correct answers.

We computed these scores for all occurrences of FEEs that have been independently annotated by two annotators so far. Results are shown in table 1. The first column provides the number of pre-annotated FEE occurrences proposed for double annotation, in total and for each domain (a FEE may belong to several domains in these counts, either because it is associated with frames from several domains, or because the frame per se pertains to several domains). The relatively low number of instances from the verbal communication domain results from this domain being only partially annotated, and containing primarily FEEs ambiguous with some of the other three domains.

Overall, agreement is roughly 86% for frames and 77% for roles of matching frames. Agreement is thus substantial, which shows that the task was well defined. The break-down by notional domain reveals that agreement varies for the different domains. It is computed for each domain, by considering only FEEs that have at least one frame belonging to the domain. Hence it mixes information concerning ambiguity both within a domain and cross-domains. Overall, the best and worst agreement for frame assignment is achieved for the commercial transaction domain and the causality domain, respectively. An explanation could be that the ambiguity within the causality domain was more difficult to cope with. As far as agreement on role assignment is concerned, the break-down by part-of-speech of the FEE seems to provide the best explanation for the agreement variation across domains: role assignment agreement is much lower for nouns than for verbs. Hence, domains with high ratio of verbs (cognitive stances, verbal communication) have a better role assigment agreement than the other two domains, which have fewer verbal FEEs. Note that the agreement achieved for verbal FEEs is on a par to that reported by Burchardt et al. (2009) for German, within the SALSA project, which focused on verbal FEEs mainly (reported agreement is 85% for frames, and 86% for roles of matching frames, whereas ours for verbs is 87.6 for frames, and 82.8%). Our having better agreement for frames may be due to our restricting ourselves to 4 domains.

[13] Note that the counts are before sense disambiguation. A given FEE is associated to a notional domain if one of its frames at least belongs to this domain. Hence the ALL count is less than the sum for all the 4 domains.

4.3. Resulting annotated corpus

We now turn to the current resulting resource. Adjudication was not fully completed at the time of writing, although it will be by the time of the conference (this touches about 15% of the FEE occurrences). All the figures provided in this section have been obtained as if adjudication had been completed, using one annotator's annotations as if they resulted from adjudication. Furthermore, in order to complete annotation of the domains, some FEEs were annotated by one expert only (totalizing about 10% of all the annotated FEE occurrences).

Statistics are provided in table 2. We have about 100 frames and over 650 distinct FEEs, with various POS, though mainly verbs and nouns. A syntactico-semantic lexicon was extracted from the annotated corpus, providing precise information on the syntactic realization patterns of the roles of each frame+FEE pair and each frame overall.

5. Handling of specific linguistic phenomena

5.1. Non-strict semantic compositionality

As stressed in (Burchardt et al., 2009), corpus-based frame annotation runs into several kinds of full or partial semantic non-compositionality, namely idioms, light verb constructions, and metaphors. We treat these as they are in the Berkeley FrameNet (Ruppenhofer et al., 2006). Idioms are identified as multi-word expressions for the most part frozen, whose meanings have to be understood as a whole. Such idioms can evoke frames pretty much like plain words can, and are thus recorded in the lexicon. For instance the idiom *mettre (qqch) sur le compte de (qqun/qqch)* (litt. *to put (sth) on the account of (sb/sth)*, meaning *to blame (sth) on (sb/sth)*) can evoke the FR_Attributing_cause frame. Note that such idiomatic FEEs, especially the verbal ones, are potentially discontinuous.

Light verb constructions (LVC) combine a predicative noun and a verb which does not have exactly the meaning it would have in the absence of the noun. As in previous FrameNet projects, the frame is evoked by the noun, and we mark the verb as support verb, in order to easily spot such constructions. For that purpose, we used specific frames for support verbs, with only one "role" , for the noun.

	Nb distinct frames	Nb distinct FEEs	Nb senses	Nb annotation sets ($\neq$ dummy frame)	% of non-dummy frames
ALL	98	662	872	12874	64.4
Commercial	19	93	103	2937	-
Causality	11	217	252	3843	-
Cognitive stances	40	283	356	4040	-
Communication	36	168	210	2631	-
N	-	207	247	4275	63.9
V	-	341	483	7087	66.6
PREP	-	25	29	530	54.5
ADV	-	23	31	320	48.9
CONJ	-	21	27	299	79.3
ADJ	-	37	42	210	53.6

Table 2: Statistics of the resulting annotated resource: number of distinct frames and FEEs, number of senses (association frame + FEE) and number of annotation sets. Last column: proportion of annotated occurrences that are not the Other_sense frame.

Finally metaphors are cases of new or not conventionalized non-literal meaning, either lexical (i.e. concerning one word only) or multi-word. Although novelty is a continuous property, we map it to a binary feature, which distinguishes lexical metaphor from polysemy, and multi-word metaphors from idioms (using Ruppenhofer et al. (2006) criteria). Given our annotation workflow, polysemic senses and idioms arising from conventionalized metaphors were entered in the lexicon before annotation time. For a (non-conventionalized) metaphorical use of a lemma, the candidate frames proposed to the annotators pertained to literal meanings only. In such cases, annotators were asked to annotate the literal meaning (from the source domain in Lakoff's view), but to flag the frame occurrence as metaphorical. Contrary to what was done in the SALSA project, we did not annotate the metaphorical meaning, because such annotation might fall outside the notional domains we targeted.

5.2. Syntactic non-locality

In Berkeley FrameNet lexicographic annotations, only those role fillers that were realized "in grammatical construction" with the FEE (Fillmore, 2007) were annotated. This includes the syntactic dependents of the FEE, but also some syntactically well-defined cases of syntactic non-locality of role fillers with respect to the FEE (raising and control, relative clauses (Ruppenhofer et al., 2006, p. 27), but also light verb constructions, copular sentences and FEEs typically modifying and thus governed by one of their role). This choice derives from the objective of obtaining relevant patterns of how semantic valents realize as syntactic valents.

Using this strategy means we sometimes do not annotate a role filler which is expressed in the sentence of the FEE but outside its locality domain. Because FrameNet annotations are also useful for obtaining lexical selectional preferences, we found it interesting to annotate even non locally realized role fillers. So for instance in *La terrible pression de la grande distribution pour* **acheter** *le plus bas possible asphyxiait les fabricants (the terrible pressure of supermarkets to* **buy** *at the lowest (possible price) was asphyxiating manufacturers)*, *manufacturers* and *supermarkets* are clearly understood and thus annotated as the Seller and the Buyer of the buying frame[14].

In the same vein, when a role filler is a non-lexical anaphor with a lexical antecedent expressed within the same sentence, we asked annotators to mark both as fillers of the same role (with flags indicating the anaphoric relation), in order to capture the syntactic regularity (when the anaphor is realized as a regular syntactic valent of the FEE) without missing a lexical role filler (the antecedent).

5.3. Null instantiation

Sometimes, a role that has been deemed conceptually necessary to a frame cannot be filled by any part of a sentence where said frame is evoked. Berkeley FrameNet keeps track of those cases of what they call *null instantiation* (NI) for they provide "lexicographically relevant information regarding omissibility conditions" (Ruppenhofer et al., 2006). A distinction is made between three types of NI. When a constituent is constructionnally omitted, as is often the case of agents in passive sentences, the role that this constituent would have filled is labeled CNI for *Constructional NI*. *Definite NI* (DNI) covers cases in which the missing role must be interpreted from the linguistic or discourse context, whereas *Indefinite NI* (INI) is used for existentially understood missing roles, which for certain verbs tend to receive a stereotypical interpretation (e.g. if someone eats, it is assumed that they eat a meal).

For the French FrameNet, we have used a different set of NI subtypes. Where FrameNet uses DNI to annotate indiscriminately missing role fillers that can be found in either the linguistic or discourse context, our own Definite NI is only used when the non-instantiated role is clearly filled by an element of the linguistic context. When the role filler can be interpreted but has not been explicitly mentioned in the few surrounding sentences, we consider the role to be an *Extra-linguistic NI* (ENI). Only the DNI cases should be looked for in surrounding sentences. All the other cases of NI, that is those in which the missing role filler cannot be interpreted, are labeled as *Unknown NI* (UNI). We have

[14] A flag is used to identify non-local instantiations of FEEs. Because these are necessarily more subject to interpretation issues, annotators were asked to favor syntactic locality whenever possible, and to only annotate doubtless non-local fillers.

decided not to include a CNI-type label, as it seemed redundant to us to make a difference between arguments lexically vs. syntaxically omitted: the information as to whether an argument's omission is constructional can be extracted from the syntax itself. To sum up, the null instantiation types distinguish whether the interpretation of the missing role can be found in the linguistic context (DNI), discourse context (ENI) or is underspecified or generic (UNI).

(1) Ils revendront plus tard.
They will-resell more late. (They will resell later.)

In the sentence (1), in which *revendront* evokes the Commerce_sell frame, *ils* fills the Seller role, and the Goods to be resold (real estate) are labeled DNI, because they are mentioned in a previous sentence. By contrast, the only thing that can be said about the Buyer in that Commerce_sell event is that they are people likely to buy real estate. Since that information is entirely entailed by the definition of the Buyer role and the filler of the Goods role, we consider the Buyer to be UNI.

5.4. Syntactic alternations

In this section, we briefly review how productive syntactic alternations are treated in the French FrameNet. By productive, we mean those applying without many lexical exceptions, given a certain canonical syntactic valency. For French these are the passive, middle, impersonal, reflexive and causative alternations.

The treatment of syntactic alternations is not described at length in FrameNet's guidelines, but can be inferred from the general objectives of FrameNet. On the one hand, all LUs evoking a given frame should share the same semantic arguments, on the other hand productive syntactic alternations should be neutralized. So only those alternations that do not modify the set of participants of a predicate should be captured within the same frame. This is the case for passive, middle and impersonal alternations: personal and impersonal uses of an intransitive verb, cf. examples (2) and (3), fall under the same frame, and so do active, passive and middle voices of a transitive verb. For the middle voice, the well-known key aspect to consider is that although the agent participant (the canonical subject) cannot be locally expressed, cf. example (4), it is necessarily understood. So middle uses of a transitive verb are represented using the same frame as the transitive verb, with a null instantiation flag for the missing participant.

(2) Des complications peuvent en résulter.
Some complications may from-it result.
(Complications may result from it)

(3) Il peut en résulter des complications.
It may from-it result some complications.
(Complications may result from it)

(4) Les crises ne s' anticipent pas (*par les traders).
The crisis NEG REF anticipate not (*by the traders).
(One doesn't anticipate crisis)

The situation is different for the causative/inchoative alternation for some transitive verbs (roughly change of state or change of position verbs). The transitive version can be understood approximately as providing a cause for the intransitive version. But unlike in the middle construction, in the intransitive version (the neuter construction) the agent needs not be understood: it is either absent or at least deprofiled. For this reason, neuter versions cannot be captured within the same frame as the transitive construction. A frame-to-frame relation is used between the frames for the transitive and the intransitive versions.

(5) Le vase peut (se) casser.
The vase can (REF) break. (The vase can break)

The causative alternation also modifies the number of participants. French causative constructions involve the *faire (to make)* verb combining with an infinitive, e.g. (6). From the semantic point of view, except for cases in which the combination of *faire* and the infinitive is lexicalized, the meaning of the construction is roughly compositional: one extra semantic argument (the causer) triggers the eventuality described by the infinitive, which can roughly be analyzed as a causal relation. From the formal point-of-view though, it is well-known that properties of causative constructions in Romance languages lead to two competing analysis (Abeillé et al., 1996), in which the combination of *faire* with the infinitive is either regular or forms a complex single predicate. In the syntactic annotation scheme of our target corpora, only the latter was retained: the *faire* verb forms a complex predicate with the infinitive verb, which appears with a non canonical valency (hence the characterization of causative constructions as syntactic alternations). The causer appears as subject of the complex predicate, and the causee is demoted to direct or oblique object, depending on the infinitive's transitivity.

(6) Il veut faire payer le surcoût à Paul.
He wants-to make pay the overcost to Paul.
(He wants to make Paul pay the overcost.)

Because the causative uses of an infinitive have an extra participant (the causer), they cannot evoke the same frame as in the non-causative versions. We annotate causative constructions compositionally, which results in a syntax/semantic divergence: *faire* evokes the Causation frame, with the causer filling either the Cause or the Actor role, and the infinitive regularly evoking one of its frames.

Finally, the (true) reflexive alternation reduces the set of participants in that two roles are played by the same entity. Furthermore in French, the reflexive clitic *se* is considered as semantically void, meaning that the reflexive verb has one fewer syntactic valent. So we simply annotate true reflexives by assigning two roles to the subject.

5.5. Classification of the reflexive *se* marker

As noticeable from the previous section, the French reflexive *se* clitic paradigm is highly ambiguous: it can mark (i) lexicalized *se*+V combinations, neuter and middle syntactic alternations (cf. examples (5) and (4) in the previous section), true reflexive or reciprocal constructions. Depending on the status of the clitic, a given occurrence of *se*+V triggers a frame evoked by the V alone, or by the lexicalized combination *se*+V. In order to break up annotation tasks,

we used a preliminary annotation of the reflexive *se*'s status in context[15]: when faced with a *se*+V occurrence, the *se*'s status was used by the pre-annotation tool to choose whether to pre-annotate frames pertaining to the V alone (in case of middle, true reflexive or reciprocal) or to the clitic+V combination.

6. Conclusion

We presented the annotation workflow for a French FrameNet, consisting of a set of frames, a lexicon, and frame and role annotations added to syntactic treebanks. The resource is focused on four notional domains (commercial transactions, cognitive stances, causality and verbal communication). Inter-annotator agreement is substantial, and the resulting resource can be used for word sense disambiguation and semantic role labeling tasks, as well as for studying syntactic valency patterns of semantic frames.

7. 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). We thank very much the people involved in the preceding frame and lexicon-building phase of the project, and are extremely grateful to the annotators for their work: Vanessa Combet, Nomie Faivre, Virginie Mouilleron, Marjorie Raufast, Anny Soubeille and Emilia Verzeni.

8. Bibliographic References

A. Abeillé and N. Barrier. 2004. Enriching a french treebank. In *Proc. of LREC'04*.

A. Abeillé, D. Godard, and P. Miller. 1996. Les causatives en français, un cas de comptition syntaxique. *Langue Française*, 115:62–74.

C. F. Baker, C. J. Fillmore, and J. B. Lowe. 1998. The Berkeley FrameNet project. In *COLING-ACL '98: Proc. of the Conference*, pages 86–90.

H. C. Boas, editor. 2009. *Multilingual FrameNets in computational lexicography : methods and applications*. Trends in linguistics. Mouton de Gruyter.

L. Borin, D. Dannlls, M. Forsberg, M. Toporowska Gronostaj, and D. Kokkinakis. 2009. Thinking Green: Toward Swedish FrameNet++. In *FrameNet Masterclass and Workshop*.

A. Burchardt, K. Erk, A. Frank, A. Kowalski, S. Padó, and M. Pinkal. 2006a. SALTO – A Versatile Multi-Level Annotation Tool. In *Proc. of LREC'06*.

A. Burchardt, K. Erk, A. Frank, A. Kowalski, S. Padó, and M. Pinkal. 2006b. The SALSA Corpus: a German corpus resource for lexical semantics. In *Proc. of LREC 2006*.

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 Boas (Boas, 2009), pages 209–244.

[15]This annotation was performed by V. Combet, V. Mouilleron, B. Sagot and M. Candito at the occasion of the ASFALDA project and the deep Sequoia project (https://deep-sequoia.inria.fr).

M. Candito and D. Seddah. 2012. Le corpus sequoia : annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. In *Proc. of TALN 2012 (in French)*.

M. Candito, P. Amsili, L. Barque, F. Benamara, Ga. de Chalendar, M. Djemaa, P. Haas, R. Huyghe, Y. Y. Mathieu, P. Muller, B. Sagot, and L. Vieu. 2014. Developing a French FrameNet: Methodology and first results. In *Proc. of LREC'14*.

D. Das, N. Schneider, D. Chen, and N. A. Smith. 2010. Probabilistic frame-semantic parsing. In *Proc. of NAACL HLT 2010*, pages 948–956.

C. J. Fillmore, 1982. *Frame semantics*, pages 111–137. Hanshin Publishing Co.

C. J. Fillmore. 2007. Valency issues in framenet. In Thomas Herbst and Katrin Gotz-Votteler, editors, *Valency: theoretical, descriptive and cognitive issues*, volume 187 of *Trends in Linguistics*, pages 129–162. Mouton de Gruyter.

B. Levin. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press.

C. Mouton, G. de Chalendar, and B. Richert. 2010. Framenet translation using bilingual dictionaries with evaluation on the English-French pair. In *Proc. of LREC'10*.

K. Ohara, S. Fujii, S. Ishizaki, T. Ohori, Hiroaki S., and R. Suzuki. 2004. The Japanese FrameNet project; an introduction. In *Proc. of the Workshop on Building Lexical Resources from Semantically Annotated Corpora*. LREC'04.

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

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

J. Ruppenhofer, M. Ellsworth, M. R.L. Petruck, C. R. Johnson, and J. Scheffczyk. 2006. *FrameNet II: Extended Theory and Practice*. International Computer Science Institute, Berkeley, California. Distributed with the FrameNet data.

K. Kipper Schuler. 2005. *Verbnet: a broad-coverage, comprehensive verb lexicon*. Ph.D. thesis.

D. Seddah, R. Tsarfaty, S. Kübler, M. Candito, J. Choi, R. Farkas, J. Foster, I. Goenaga, K. Gojenola, Y. Goldberg, S. Green, N. Habash, M. Kuhlmann, W. Maier, J. Nivre, A. Przepiorkowski, R. Roth, W. Seeker, Y. Versley, V. Vincze, M. Woliński, A. Wróblewska, and E. Villemonte de la Clérgerie. 2013. Overview of the SPMRL 2013 Shared Task: A Cross-Framework Evaluation of Parsing Morphologically Rich Languages. In *Proc. of the 4th Workshop on Statistical Parsing of Morphologically Rich Languages: Shared Task*.

C. Subirats-Rüggeberg and M. R.L. Petruck. 2003. Surprise: Spanish FrameNet! In *Proc. of the Workshop on Frame Semantics, XVII International Congress of Linguists (CIL)*. Matfyzpress.