

## VARIATION IN VERBAL PREDICATES IN ENGLISH AND FRENCH

Tanja Samardzic (*Tanja.Samardzic@unige.ch*), Lonneke van der Plas (*Lonneke.vanderPlas@unige.ch*), Goljihan Kashaeva (*Goljihan.Kashaeva@unige.ch*), Paola Merlo (*Paola.Merlo@unige.ch*)

### 1. INTRODUCTION

The analysis of the predicate-argument structure of a sentence results in a linguistic representation that defines relations between the constituents in the sentence that cannot be defined by the rules of syntax only. Different kinds of syntactic subjects, such as those in (1b) and (1c) for example, can only be accounted for at this level of representation. Moreover, this level of analysis provides a means of expressing a relation between syntactically different sentences such as the sentence with the transitive verb in (1a) and the one with the intransitive verb in (1b). The semantic label *THEME* expresses the fact that the object in (1a) stands in the same relation to the verb as the subject in (1b).

- (1) a. [AGENT Mary] stopped [THEME the car].  
b. [THEME The car] stopped.  
c. [AGENT Mary] stopped.

Cross-linguistically, the predicate-argument structure of a sentence is considered to be more stable (less varying) than its syntactic form. The English sentence in (2a) can be considered as equivalent to the French sentence in (2b), despite the fact that the positions of their syntactic subjects are occupied by different kinds of lexical elements and that the complements of the verbs differ both syntactically and semantically. The predicate-argument structure of (2a) and (2b) is equivalent since verbs (*liked*, *a plu*), which are the predicates of the sentences, take the same kind of arguments (*EXPERIENCER*, *CONTENT*) in both languages.

- (2) a. [EXPERIENCER Mary] liked [CONTENT the idea]. (English)  
b. [CONTENT L'idée] a plu [EXPERIENCER à Marie]. (French)

Not all predicates behave in the same way in a cross-linguistic context. Some of them are known to give rise to more parallel or cross-linguistically stable syntactic structures, while others result in more divergent syntactic realizations. For instance, the predicates expressed by verbs meaning “creation”, such as English *create*, *build*, *construct* are likely to be realized as transitive verbs with direct objects in many different languages, while predicates denoting some state of mind, such as the verbs in (2), can be involved in different syntactic structures in different languages (Levin & Rappaport Hovav 2005).

Linguistic investigations into predicate-argument structure have been mostly concerned with the kinds of labels (semantic or thematic roles) that can be assigned to arguments. The labels that are most commonly used can be related to the “deep cases” in Case grammar (Fillmore 1968). Beside those in (1), the set of traditionally and most widely used labels includes: *EXPERIENCER*, *INSTRUMENT*, *SOURCE*, *GOAL*, *LOCATION*. More recently, a number of approaches have been proposed to systematise the traditional inventory of roles and to relate the number and the labels of arguments to different semantic properties of the predicates (Dowty 1991, Reinhart 2002, Jackendoff 1990, Levin & Rappaport Hovav 2005). These

approaches typically give up the traditional notions and adopt abstract roles defined in terms of sets of features (Dowty 1991, Reinhart 2002) or in terms of different components of the meaning of the predicate (Jackendoff 1990).

In addressing cross-linguistic variation, these approaches assume that the same predicates are associated with the same argument structures in different languages. The focus is on the differences in syntactic realisation of the arguments and the semantic features that give rise to the variation. Little attention has been given to cross-linguistic variation in the argument structure itself, that is, to the potential differences in the number and the kind of arguments assigned by the predicates. However, a corpus-based analysis of the variation between English and Chinese (Fung et al. 2007) suggests that cross-linguistic parallelism in the predicate argument structure cannot be assumed. This study shows that 17% of arguments of English predicates do not map to the arguments of the corresponding Chinese predicates.

The aim of our study is to examine the variation of the predicate-argument structure between English and French predicates in an experimental and data-driven approach. We attempt to annotate the predicate-argument structure in a corpus of one-thousand naturally occurring French sentences using an annotation framework previously developed for English. We identify and analyse a number of French predicates which are not accounted for in the English framework and, thus can be considered as varying. Such an approach rests on two important methodological decisions.

First, we compare two closely related and well-documented languages. This choice is based on the assumption that English and French constitute a minimal pair suitable for micro-comparisons (Kayne 2005). Since the variation in the predicate-argument structure is related to lexical items rather than to broader syntactic structures, micro-comparison seems to be the adequate approach, providing a good setting for identifying the elements of grammar that potentially underlie the variation.

Second, we take a corpus-based approach, analysing a predefined sample of naturally occurring sentences and all verbal predicates that occur within these sentences. The motivation for this approach comes from the fact that it enables collecting and processing large amounts of data required for studying linguistic variation. Furthermore, an exhaustive analysis of a limited sample of sentences provides a means for assessing the scope of variation as a proportion of units affected by it.

We argue that the two languages do not differ substantially in the inventory and the nature of verbal predicates, even though certain general grammatical properties may result in some variation at this level of representation. We also argue that the proposed comparative study can provide a basis for identifying the level of specificity needed for developing a framework for multilingual annotation of predicate-argument structure.

The issues addressed in this article are of interest both for theoretical and computational linguistics. On the one hand, we are making use of the available computational resources and tools for addressing linguistic issues that require a large scale data analysis. On the other hand, the results of such an analysis can be directly incorporated in developing or improving tools for automatic natural language processing, especially automatic analysis of predicate-argument structure.

A successful automatic analysis of the predicate-argument structure of a sentence requires an approach that combines deep linguistic analysis with the methods developed in natural language processing. It makes use of extensive linguistic resources, corpora with annotated instances of predicates and semantic roles of their arguments. Developing such corpora requires linguistic expertise, since the labels used in the annotation necessarily depend on a theoretical framework and the notions developed within the framework. In the other direction, annotating a corpus in a theoretical framework can be seen as a test of the

descriptive adequacy of the framework, since the concepts developed as part of this framework need to be applied to analyse a great number of different cases.

Developing linguistic resources with annotated predicate-argument structure is a long and expensive process. For the moment, the only originally developed annotation frameworks are those developed for annotation of English corpora. There is great interest in investigating cross-linguistic adequacy of these frameworks and their applicability to other languages. Here again, the interest is two-fold. On the side of computational linguistics, direct application of a framework developed for one language to other languages would make it possible to quickly develop systems for automatic analysis of the predicate-argument structure in many different languages. On the side of theoretical linguistics, it is desirable for a model of the structure to be applicable to multiple languages. Its cross-linguistic validity depends not only on the degree of variation between the languages, but also on the use of language-specific notions. A comparative study can help identify and eliminate elements of the framework which are too language specific.

The article is organised as follows. In Section 2, we discuss the existing methods for large-scale analyses and corpus annotation of the predicate-argument structure. We briefly review the frameworks and the resources developed for English, with a special reference to PropBank, the resource we choose for our research. We give the details on our method in Section 3. In Section 4, we describe the varying predicates and propose four types of cross-linguistic divergence. In Section 5, we propose a unifying analysis of the divergent structures identifying the components of the structures responsible for the observed variation. Finally, we draw some conclusions in Section 6.

## 2. ANNOTATION FRAMEWORK

There are three frameworks proposed for annotating corpora with predicate-argument structure: FrameNet, VerbNet and PropBank. All of them have been developed on the basis of English data and have been used to annotate English corpora. However, these frameworks are implementations of linguistic theories of predicate argument structure that have been developed to account for universal phenomena, so they can be expected to apply to other languages as well.

In the following subsections, we describe the main characteristics of the three annotation frameworks and the criteria we used in choosing the PropBank annotation for our study.

### 2.1. FrameNet

FrameNet is a resource intended to be used by lexicographers, but also by systems for natural languages processing (Baker et al. 1998). It consists of two interrelated databases.

The *frame database* (946 frames) contains descriptions of frames or scenes that can be denoted by predicating lexical units (Fillmore 1982), such as verbs, adjectives, prepositions, nouns. Each scene involves one or more participants that are denoted by the lexical units which combine with the predicating units. The predicating units are referred to as “targets”, and the semantic roles that they assign as “frame elements”. One frame can be realized in its “core” version including “core” frame elements, or it can be realized as a particular variation, including additional frame elements that are specific for the variation. For example, the target unit for the frame *Accomplishment* can be one of the verbs *accomplish*, *achieve*, *bring about*, or one of the nouns *accomplishment*, *achievement*. The core frame elements for this frame are AGENT (the conscious entity, generally a person, that performs the intentional act that fulfills the goal) and GOAL (the state or action that the agent has wished to participate in), while

different versions can include additional frame elements as in (3). Each frame introduces a different set of role labels, which can be general, such as those used in (3), or rather specific such as BUYER for the frame *Commerce\_buy*.

- (3) a. [AGENT Iraq] had [TARGET achieved] [GOAL its programme objective of producing nuclear weapons].  
 b. Perhaps [AGENT you] [TARGET achieved] [GOAL perfection] [MANNER to quickly].  
 c. [AGENT He] has [DEGREE only partially] [TARGET achieved] [GOAL his objective].

The *lexicon* contains 5161 fully described and 5560 partially described lexical units (word senses). The descriptions specify the frame that the unit realizes, the list of the frame elements that can occur with it, and the syntactic form that each frame element can take. For example, the verb *achieve* realizes the frame *Accomplishment*. The frame elements that can occur with it are listed in Table 1.

Frame element	Syntactic realization (phrase types and their functions)
agent	CNI.– NP.Ext PP[by].Dep PP[for].Dep
circumstances	PP[in].Dep PP[despite].Dep PP[as].Dep AJP.Dep PP[on].Dep
degree	AVP.Dep
explanation	PP[since].Dep PP[because of ].Dep
goal	NP.Ext NP.Obj NP.Dep PP[in].Dep
instrument	PP[with].Dep NP.Ext VPing.Dep
manner	AVP.Dep PP[in].Dep
means	PP[through].Dep PP[by].Dep PP[in].Dep NP.Ext
outcome	PP[at].Dep NP.Ext
place	PP[in].Dep PP[at].Dep
time	Sfin.Dep AVP.Dep PP[in].Dep PP[at].Dep PP[after].Dep

Table 1: Frame elements listed in FrameNet for the verb *achieve*.

The first row in Table 1 states that the frame element AGENT occurs with the verb *achieve* and that it can be realized as Constructional null instantiation (CNI), which is most often the case in passive sentences. It can also be realised as a noun phrase external to the verb phrase headed by the target verb, which is most often the subject of a sentence, or as a prepositional phrase headed by the preposition *by* and realizing the grammatical function of dependent<sup>1</sup> or as a prepositional phrase headed by the preposition *for* with the same grammatical function. Possible syntactic realizations for the other frame elements are described in the same way.

<sup>1</sup> In the system of grammatical functions used in FrameNet, the standard distinction between a complement and a modifier is not made. They are both considered as dependent constituents – dependents. (Ruppenhofer et al. 2005)

In addition to these two databases, FrameNet contains a corpus of manually annotated sentences selected from the British National Corpus (Burnard 2007) as examples for most of the frame versions.

## 2.2. VerbNet

VerbNet is primarily concerned with the classification of English verbs (Kipper 2005). It describes 5200 verb senses (of 3600 distinct verbs). The approach to classification is based on the framework proposed by Levin (1993). It takes into account two properties: a) the lexical meaning of a verb and b) the kind of argument alternations that can be observed in the sentences formed with a particular verb. The idea that underlies Levin's classification is that syntactic constraints reflect certain semantic properties of verbs. Thus, observing syntactic behaviour of verbs could provide the background for their systematic and consistent classification that holds across languages.

The verbs are classified into 237 classes, where all the verbs belonging to the same class assign the same roles to their arguments. The label for a role in VerbNet does not depend on the predicate as in FrameNet. There is a fixed set of 23 roles that have the same interpretation with all verbs: ACTOR, AGENT, ASSET, ATTRIBUTE, BENEFICIARY, CAUSE, LOCATION, DESTINATION, SOURCE, EXPERIENCER, EXTENT, INSTRUMENT, MATERIAL, PRODUCT, PATIENT, PREDICATE, RECIPIENT, STIMULUS, THEME, TIME, TOPIC.

## 2.3. The Proposition Bank

The Proposition Bank (Palmer 2005) is a corpus (one million words) of naturally occurring sentences with manually annotated predicate-argument structure. It is intended to be used for developing systems for semantic parsing, but also for quantitative analysis of syntactic alternations and transformations.

Only a limited set of labels based on Dowty's theory of Proto-Roles (Dowty 1991) was used for annotation. Verbs are marked with the label REL for relation and the participants in the situation denoted by the verb are marked with the labels ARG0 to ARG5 for the verb's arguments and with ARG-M for adjuncts.

The numbered labels represent semantic roles of a very general kind. The labels ARG0 and ARG1 have approximately the same value with all verbs. They are used to mark instances of PROTO-AGENT (ARG0) and PROTO-PATIENT (ARG1). The meaning of other numbered arguments varies across verbs. It depends on the meaning of the verb, on the type of the constituent that they are assigned to, and on the number of roles present in a particular sentence. ARG3, for example, can mark PURPOSE, as it is the case in (4), or it can mark DIRECTION or some other role with other verbs. The indices are assigned according to the roles' prominence in the sentence. More prominent are the roles that are more closely related to the verb. The ARG-M labels can be specified further as: LOCATION, CAUSE, EXTENT, TIME, DISCOURSE CONNECTIVES, PURPOSE, GENERAL PURPOSE, MANNER, DIRECTION. The labels for adjuncts are more specific than the labels for arguments. They do not depend on the presence of other roles in the sentence.

- (4) [ARG0 The Latin American nation] has [REL-PAY.01 paid] [ARG1 very little]  
[ARG3 on its debt] [ARG-M-TMP since early last year].

The annotated corpus is accompanied with a set of *frame files* that specify the interpretation of the roles for each verb in its different senses. Each lemma described in the frame files (3300 verbs) contains one or more lexemes (4 500 verb senses), which are used as predicate

labels (Frames in Table 2). The interpretations for the numbered roles are given for each lexeme separately. Table 2 illustrates the lexical entry for the verb *pay*.

Frame	Semantic roles
pay.01	ARG0: payer or buyer ARG1: money or attention ARG2: person being paid, destination of attention ARG3: commodity, paid for what
pay.02 pay off	ARG0: payer ARG1: debt ARG2: owed to whom, person paid
pay.03 pay out	ARG0: payer or buyer ARG1: money or attention ARG2: person being paid, destination of attention ARG3: commodity, paid for what
pay.04	ARG1: thing succeeding or working out
pay.05 pay off	ARG1: thing succeeding or working out
pay.06 pay down	ARG0: payer ARG1: debt

Table 2 : The PropBank lexicon entry for *pay*.

## 2.4. Choosing the framework

All three resources describe the same linguistic phenomena covering approximately the same range of lexical items. Nevertheless, they are conceived with different purposes and with different theoretical background. As a result, the choice of the framework can have important consequences for the outcome of the annotation, especially in a cross-linguistic setting.

Predicate labels used in FrameNet (frames) and VerbNet (verb classes) are intended to capture the level of lexical semantics which is common to a group of lexical items. Different words can bear the same label. For instance, the verbs *accomplish*, *achieve*, *bring about*, already mentioned in 2.1, all bear the label Accomplishment in FrameNet. This is not the case in the PropBank framework, where predicate labels capture the specific meaning of verb senses. This is why the style of annotation used in FrameNet and VerbNet can be considered as more abstract and, thus, more portable across languages than the annotation in PropBank. Indeed, FrameNet has often been used as a basis for developing similar resources for other languages (Burchardt 2009). Furthermore, argument labels in FrameNet are assigned without taking into account the syntactic function of the constituents that bear them, while the PropBank argument labels can depend on the syntactic function, especially those with higher indices (ARG2-5). This feature renders the FrameNet annotation less tied to the syntactic representation and thus to a particular language than it is the case with the PropBank annotation.

Although PropBank is considered as the most language specific of the three resources, we use this framework to annotate the predicate-argument structure of French sentences. We take this decision for two reasons. First, the lexicon in this resource is corpus-driven. It is built by extracting and describing all the predicates that occur in a predefined sample of

naturally occurring sentences. Since our aim is to annotate exhaustively a corpus of naturally occurring sentences, we can expect that such a lexicon can provide a better coverage than the lexicon in FrameNet and VerbNet, which are not corpus-driven. Second, the labels used in PropBank both for predicates and arguments involve fewer theoretical assumptions than the labels in FrameNet. While the FrameNet and labels capture mostly linguistic intuition at the targeted level of lexical semantics and the relations between the lexical items, the PropBank labels rely strongly on the observable behaviour of words. The distinction between the different verb senses, for instance, is made taking into account the different sets of arguments and other observable differences, such as the presence of the particle that distinguishes *pay.05* from *pay.04* in Table 2. This approach can be expected to provide more tangible criteria for annotators in deciding how to annotate each instance of the predicate-argument structure found in the corpus, ensuring a more reliable and more consistent annotation. Also, it enables a more direct comparison of the structures across languages, since the representation of the structures does not include any hypothesized levels of abstraction.

### 3. MATERIALS AND METHODS

In identifying the points of variation in the predicate-argument structure between English and French, we rely on corpus data. We manually annotate a corpus of 1000 French sentences using the PropBank annotation framework originally developed for English. Those French predicates for which no appropriate PropBank label (English verb sense) can be found are considered as varying. The sentences containing these varying predicates were automatically extracted from the corpus and manually analysed.

We normalized the instances of predicates extracted to their citation form and looked up a comprehensive French-English dictionary to identify the English translation of the French predicate that is closest to the literal translation. If needed, the translations were verified with a native speaker. For example, we find the English expression *shed light* as the translation for the French expression *faire la lumière*. The list of all extracted instances can be found in the Appendix. After having applied this procedure to all varying predicates extracted from the corpus, we classify and analyse the predicates. We discuss the proposed classification in Section 4.

We envisage to detect cross-linguistic variation by looking at those cases where the manual annotation using a resource developed for a source language (English) fail to be applicable to the target language (French). It must therefore be ensured that failure is due to a genuine inability to provide a common analysis for the two languages and not to other causes. To ensure good quality and reliability of the annotation, as well as of the identification of non-matching predicates, we provide the annotators with detailed guidelines and set up two training phases and a calibration phase. The following two subsections describe these steps in more detail. The corpus of French sentences is described in 3.3.

#### 3.1. Annotation tool and guidelines

We provide the annotators with guidelines that consist of an extensive description of the PropBank annotation framework (adapted from the original PropBank guidelines (Babko-Malaya 2005)), parallel English and French examples of predicate-argument analysis, and a manual on how to use the annotation tool.

The annotation tool we use is an adaptation of the user-friendly, freely available Tree Editor (TrEd, Pajas and Stepanek 2008). In Figure 1 we can see a screenshot of the tool. The tool shows the syntactic analysis and the plain sentence in the same window allowing the user to add semantic arcs and labels to the nodes in the syntactic dependency tree.

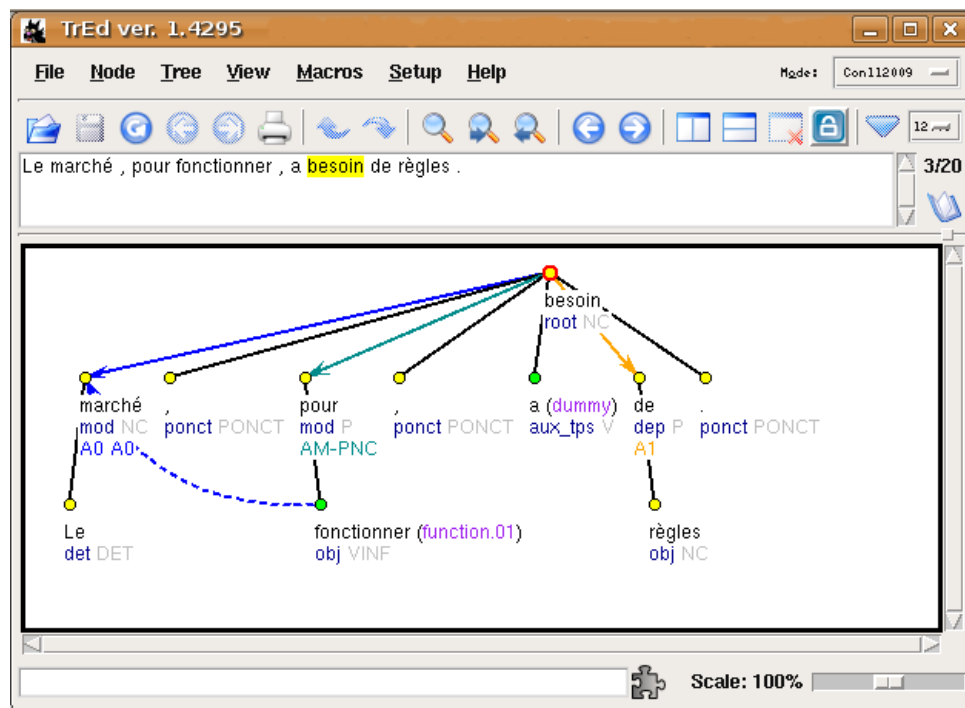


Figure 1: The TrEd annotation tool

We show syntactic information to guide the annotator in selecting the heads of phrases during the annotation process. The sentences are parsed by an automatic syntactic parser (Titov & Henderson 2007) that we trained on syntactic dependency annotations for French (Candito et al. 2009). In case of parse errors (the parser's Labelled Attachment Score is 87.2%), we ask annotators to ignore the errors of the parser and put the label on the actual head.

The guidelines provide the annotators with the following stepwise procedure for the annotation of each sentence, repeated for all verbs in the sentence:

#### Step 1: Find the predicate.

The annotator is instructed to look for a verb which heads a clause or an adjective which is the past participle form of a verb, except the verb *être*, which is not annotated in accordance with the PropBank annotation guidelines for English.

#### Step 2: Find the arguments of the predicate

These are the words or phrases that denote the semantic elements related by the predicate, usually the subject of the clause and the complements of the verb.

#### Step3: Find the corresponding frame in the PropBank lexicon

For every predicate they find in the French sentence, the annotators need to translate it to English and find the appropriate frame (English verb sense) in the PropBank frame files. They need to choose the entry that contains the frame with the meaning most similar to the meaning of the French predicate and with the set of semantic roles most similar to the set found in the French sentence. If an appropriate entry cannot be found in the frame files for a given predicate, the annotators are instructed to label the predicate as "dummy" and fill in the roles according to their own insight.

#### Step 4: Label the predicate with the frame labels and its arguments with the corresponding semantic role labels (A0-5)



Here the annotator needs to make use of the syntactic analysis presented in the lower part of the interface and insert the semantic labels in the syntactic tree. If a semantic role spans over a phrase (and not just a single word), the annotator is instructed to attach the label to the syntactic head of the phrase.

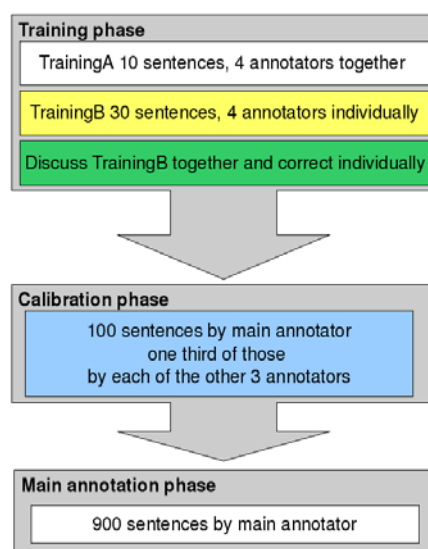
**Step 5: Find the adjuncts and label them with the corresponding AM labels.**

If an adjunct is a global modifier (modifying the whole sentence and not the particular predicate), the annotator is instructed to relate it to the highest verb in the tree.

The guidelines discuss the treatment of eight special cases: clauses as arguments, unexpressed arguments, arguments with co-reference, coordinated arguments, discontinuous arguments, expletives, fixed expressions, and English quasi-synonyms.

### 3.2. The reliability of the annotations

The main annotation was preceded by two training phases and a calibration phase. Figure 2 shows the flow of annotation phases, inspired by the methodology indicated in Pado (2007).



*Figure 2: The training phases*

During the training phase A, the annotators (three trained linguists with a very good proficiency in both French and English) and the instructor annotated 10 sentences together. In the training phase B, all annotators annotated 30 sentences. The sentences were first annotated individually. Then the annotators met to compare the annotation and discuss difficulties and differences. After discussion the annotators re-annotated the sentences trying to conform to the agreed criteria. The inter-annotator agreement was measured before and after the discussion. In the calibration phase, the main annotator annotated 100 sentences. These same 100 sentences were split up in equal portions to be annotated by the other annotators. We then measure the inter-annotator agreement for the calibration phase. In the main annotation phase the main annotator annotates the remaining 900 sentences. The guidelines were adjusted after the training phase.

Figure 3 shows the inter-annotator agreement (measured as average F-score) at different points of the annotation process for predicates and arguments separately. The first bar shows the results after individual annotation in training B. We can see that the agreement is low. The second bar shows the results after discussions which are much better. The scores are between 91% and 95% after corrections were made. This indicates that many differences between the annotators could be resolved in the discussion. This means that the task is well-defined and that a single solution was found.

The last bar shows that the agreement in the calibration phase increases significantly compared to the individual annotation in training B (before the discussion). This means that the agreement reached in the discussion could be extended to new instances which were not discussed between the annotators.

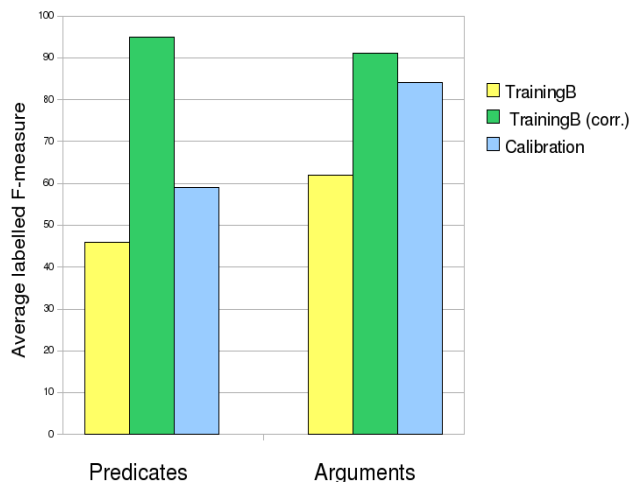


Figure 3: Inter-annotator agreement on predicate and argument labelling

The task of labelling predicates proved more difficult than labelling semantic roles of the arguments. It results in lower agreement scores. This could be expected since labelling of each French predicate requires finding the precise English verb sense that corresponds to it. Even though the verb senses determined as predicate labels in PropBank are coarser than, for example, the verb senses in WordNet (Felbaum 1998), these labels capture rather specific verb meaning. The annotators have to decide, for example, whether to use a frame of English *think* or *believe* to label an instance of French *penser*. In such cases, the agreement score is penalized by fine lexical variation in translations, which are not necessarily related to the variation in the predicate-argument structure.

This is why we measure the agreement at a more abstract level, at the level of VerbNet verb classes. We choose VerbNet since it is compatible with PropBank. The classes of verbs defined in VerbNet group together the verbs which not only have a common meaning, but also participate in the same argument alternations, exhibiting the same syntactic behaviour. For example, the verbs *think* and *believe* are in the same class, together with other epistemic verbs. We use the mapping of the PropBank labels to the VerbNet classes provided in the type mappings of the SemLink project (Loper et al. 2007). If two annotators used two different predicate labels to annotate the same verb, but those verb senses belong to the same verb class, we count those as correct. The mappings from PropBank verb sense labels to VerbNet verb classes are one-to-many and not complete. We counted a pair as matching if there exists a class to which both verb senses belong. We found a verb class for both verb senses in about 78% of the cases and discarded the rest. In Figure 4 we show that, at the level of verb classes, the inter-annotator agreement increases to 81% in the calibration phase.

The inter-annotator agreement analysis shows that using English labels to annotate French predicates is a difficult task. However, it is well defined and feasible. The improvement of the agreement at different stages of training and in the calibration phase shows that the training improves the quality and reliability of the annotation. Finally, the agreement score obtained using VerbNet class labels is very good, meaning that the annotation is reliable.

### 3.3. Corpus

The French corpus that we annotated consists of 1040 sentences drawn from the French portion of the Europarl corpus (Koehn 2005). This corpus contains translations of the proceedings of the European Parliament in 11 languages parallelized at the sentence level. We use 40 sentences for the training, so that the final French corpus annotated with predicates and arguments contains 1000 sentences.

Even though we use the French side as a monolingual corpus in our study (the annotators do not have access to the actual translations), we take the sample from a parallel corpus because this gives us a possibility to look up the actual translations of the expressions we are interested in while analysing the variation. In addition, manual annotation of one side of a parallel corpus can be a useful resource for experiments in automatic annotation transfer.

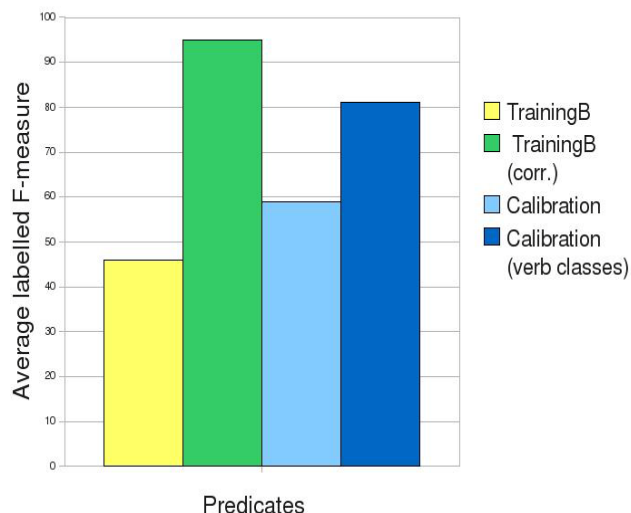


Figure 4: Inter-annotator agreement on predicates including verb class information

## 4. NON-MATCHING PREDICATES

As already described in Section 3, we instructed the annotators to use the label "dummy" if they could not find an appropriate predicate label for a French predicate in the English PropBank lexicon. Given the training that the annotator has received (see 3.1 and 3.2), as well as the fact that the task is well defined (inter-annotator agreement score of 95% after discussions), we can consider these labels as indicators of the cases where the predicate-argument structure differs between the two languages.

### 4.1. The rate of divergent structures

We find 90 instances of dummy labels for a total of 1985 predicates in the 1000 sentences of the annotated corpus. The ratio of fewer than 5% of mismatching predicates indicates that a great majority of French predicates found in our corpus directly corresponds to an English verb sense with the same predicate-argument structure. The PropBank lexicon provides a better coverage for our corpus than the FrameNet lexicon for annotating a German corpus of a similar size (Burchardt et al. 2009), where 30% of instances of German predicates required adding a label to the existing FrameNet set. The ratio of the non-matching cases that we find is also much lower than the one found in (Fung et al. 2007) for English and Chinese. According to Fung et al. (2007), 17% of arguments of Chinese verbs do not correspond to the arguments of their English counterparts. The greater variation found by Fung et al. (2007) could be due to the fact that they count mismatches in verb instances, while we are concerned with verb types. It can also be due to the fact that they are comparing distant languages, while we are comparing closely related languages.

In the following sections, we group and analyse the non-matching predicates in an attempt to identify the categories that are potential sources of variation. Our analysis does not

address idioms (3 out of the 90 instances). We also exclude one mismatch which is clearly not due to the linguistic reasons (the PropBank lexicon did not contain the verb *dramatise* for the French verb *dramatiser*). This leaves us with 86 instances that are discussed.

#### 4.2. The degrees of parallelism in mismatching predicates

The degree of parallelism between the corresponding expressions in the two languages is defined in terms of the cross-linguistic stability of the lexical categories and syntactic properties of the lexical items involved in the expressions. On the basis of the assumption that the lexical category is not an inherent feature of a lexical item (Hale & Keyser 1993), we consider as matching lexical items such as the English verb *need* and the French noun *besoin*, even though they are associated with different lexical categories in the two languages. By syntactic properties, we understand the subcategorisation frame of a lexical item, specifying the number and form of its complements.

The French structures that are most parallel with English are those where the corresponding lexical item preserves its category across languages, and the arguments of the predicates preserve their syntactic form (Group 1). Expressions that include predicates that require different syntactic forms of their arguments in the two languages are considered less parallel (Group 2). Group 3 includes the expressions where both the category of corresponding lexical items differs and syntactic divergences take place. Lastly, for a number of French predicates and their English counterparts it was not possible to find any lexical correspondence (Group 4).

##### 4.2.1. Group 1

The structures that are most parallel with English are those where the corresponding lexical item preserves its category across languages, and the arguments of the predicates preserve their syntactic form. This case is illustrated in (4). The predicates belonging to this group, as well as their English equivalents are listed in (5).

The predicates in the two sentences are realised with non-parallel lexical items. A literal English translation of French *tirer* would be *draw*, and not *learn*, which is required in this context. This predicate has received the "dummy" label due to the lexical differences between the verbs. However, the arguments of the predicate have parallel forms in the two languages.

- (4) a. Nous tirons les **leçons** du passé. (French)  
 b. We learn the **lessons** of the past. (English)

- (5) avoir lieu : take place;  
 avoir rien à voir : have nothing to do;  
 attirer l'attention : draw attention;  
 céder la place : make room;  
 faire la lumière : shed light;  
 faire alliance : form an alliance;  
 faire pression : put pressure;  
 faire appel : make an appeal;  
 porter un jugement : make an assessment;  
 tirer la leçon: learn the lesson.

The parallel structures such as these are impossible in the constructions with intransitive verbs, since they do not include the direct object which provides the lexical parallelism in transitive constructions.

#### 4.2.2. Group 2

A case of a stronger mismatch is illustrated in (6). The different verbs (predicates) used in French and English require slightly different syntactic forms of their arguments, which can otherwise be considered as direct lexical translations. The embedded clause *de réduire le niveau global des aides* is the complement of the noun *objectif* in the verb phrase, while the corresponding English clause *to reduce the overall level of aid* is the subject of the corresponding predicate (copular *be*). The lexical items *objectif* in French and *objective* in English, which head the corresponding arguments of the predicates, are direct lexical translations, but have different syntactic functions. The French argument is realised as the subject complement, while the English one is the subject of the verb. Finally, the noun *Union* is an argument and the subject of the French verb, while, in English, it is contained in the subject noun phrase, being an argument of the noun.

- (6) a. L' **Union** ne peut pas **avoir** comme **objectif** principal de **réduire** le niveau global des aides.  
 b. The main **objective** of the **Union** cannot **be to reduce** the overall level of aid.

This group includes the predicates listed in (7).

- (7) avoir qqch comme objectif : be the objective;  
 donner à qqn. à penser : make sbd. think;  
 faire l'objet de : be the objective;  
 laisser qqn. sceptique : be sceptical;  
 rediscuter : discuss again  
 tenir compte de : take into account.

The only example of syntactic divergence in the constructions with intransitive verbs is given in (8).

- (8) a. Je **m'exprime** sur le texte de M. L. (French)  
 b. I **express my opinion** on the Mr L.'s document. (English)

The pronominal verb *s'exprimer* in the French sentence can be translated with the English transitive verb *express*, with the object of the English verb corresponding to the pronominal clitic in the French verb.

#### 4.2.3. Group 3

The next degree of mismatch involves changes in the lexical category of the predicates that can be considered as lexically parallel, which also brings about syntactic divergences. A typical example of this case is given in (9) and all the predicates assigned to this group in (11). The French noun *besoin* corresponds to the English verb *need*. While the French noun is an argument of the verb *a*, the English verb is the predicate itself.

- (9) a. Le marché **a besoin de** règles. (French)  
 b. The market **needs** rules. (English)
- (10) a. Il **suffit** de lire le programme. (French)  
 b. It is **enough** to read the manifesto. (English)

The example of a stronger mismatch in intransitive verbs involve the correspondence between the French verb (*suffit* in (10a)) and an English word with a different category (the adjective *enough* in (10b)).

- (11) Transitive:  
 avoir besoin de : need;  
 avoir trait de : have to do;  
 avoir retard : be late;  
 donner suite à : follow up;  
 y faire face : face;  
 faire état : state;  
 faire défaut : lack;  
 mettre qqch en exergue : highlight;  
 porter remède : remedy;  
 prendre conscience de : be/become aware;  
 rendre compte : be accountable.
- Intransitive:  
 se mettre d'accord : agree;  
 se réjouir : be happy;  
 il suffit : be enough.

#### 4.2.4. Group 4

For a number of French predicates it was not possible to identify any lexical correspondence. A typical example of such divergence is given in (12).

- (12) a. Les dirigeants politiques doivent **faire preuve** d'un réel courage. (French)  
 b. Political leaders must **show** real courage. (English)
- (13) a. Il **s'agit** là de questions importantes. (French)  
 b. They **are** important issues. (English)

As it is the case with the predicates of the group 3 (examples (9-10)), the French verb *faire* in (12) has no corresponding English verb. However, unlike the group 3, the noun *preuve* cannot be lexically related with the English verb *show*. Similarly, the French intransitive impersonal verb *il s'agit* in (13) is translated with the English copular *are*, with the syntactic changes which necessary follow from different syntactic properties of the two verbs.

Verbs belonging to this class are listed in (14).

- (14) Transitive:  
 assigner qqn. à résidence : put sbd. under house arrest;  
 faire preuve de : show;  
 se faire un plaisir : be happy;  
 mettre qqch en oeuvre : implement;  
 metre qqch. en cause : call stg. into question;  
 remettre qqch en état : repair;  
 porter atteinte : undermine;  
 prendre la parole : speak;  
 prendre la peine : bother.

Intransitive  
 il s'agit de : be about;  
 il convient : should;  
 se féliciter : be pleased;  
 il se peut que : it is possible that;  
 se prononcer : give an opinion;  
 se rendre à : go to;  
 tenir de : be;  
 tomber bien : come in a good moment.

### 4.3. Properties of the non-matching predicates

The mismatches are mostly conventionalised expressions with different degrees of semantic compositionality, including light verb constructions and collocations. It can be noticed, though, that these expressions are not entirely idiosyncratic, but that they share some common properties.

The first property which can be noticed about the non-parallel transitive constructions is that they tend to be headed by very frequent verbs with rather general meaning. The most characteristic verb for these constructions is the verb *faire* which occurs in all four types of mismatches. It can have two kinds of meaning. One of its meanings can be described as vaguely causative, denoting that its argument which is syntactic subject (ARG0) *brings about* the entity denoted by its object (ARG1) (expressions with *faire* in (5)). The other meaning is copular, denoting that its argument which is its syntactic object (ARG1) denotes some property of the other argument (ARG0) (the expression in (7)). The meaning of the other verbs used in these constructions can be described in these terms as well. The vaguely causative meaning can be assigned to all the verbs except *avoir*, which can have a copular meaning too.

The arguments considered as ARG1 in these constructions are headed by abstract nouns, including deverbal nouns such as *pression*, *appel*, *compte*, *suite*, and *atteinte*. The meaning of these nouns is generalized with no specific reference, which is often reflected in the fact that they are not preceded by an article. All the described properties apply to all the types or degrees of non-parallelism. No characteristics specific for different groups could be observed.

There are two dominant ways in which French transitive constructions are transformed into the corresponding English expressions. The transitive verbs of the groups 1 and 2, where the lexical categories are preserved, correspond either to an English verb with the same properties (general meaning of the same type), or to the copular *be*. Transitive expressions of the group 3 are typically transformed into a single English verb, if they involve a verb with

vaguely causative meaning. Otherwise, they are transformed into a copular construction. Expressions of the group 4 are mainly transformed into a single English verb, with two examples that correspond to different English constructions and one that corresponds to the copular construction.

Expressions with intransitive verbs as predicates are less numerous than those with transitive verbs. The examples found in our corpus provide a basis for identifying two factors that can make intransitive verbs in French hard to match with English verbs. We note that most of the examples are pronominal verbs (*s'exprimer*, *se mettre d'accord*, *se réjouir*, *se féliciter*, *se prononcer*), where the reflexive clitic does not bear a semantic role. Another group of predicates are impersonal verbs (*il suffit*, *il convient*), which do not assign a semantic role to their syntactic subject and whose use is limited to a single morphological form (3<sup>rd</sup> person singular). Some predicates are characterized by both features (*il s'agit de*, *il se peut que*).

The non-parallelism in pronominal verbs can be expected, since English does not have pronominal verb forms that are parallel with French. Some kind of transformation is necessary to match these predicates. We note that they are most often transformed into a copular construction in English. This is the case with the impersonal forms too, even though they do exist for other verbs in English.

There are two expressions with intransitive verbs that are neither pronominal nor impersonal (*tenir de*, *tomber bien*). Their transformation to English forms follows the patterns identified with other predicates. The expression *tenir de* corresponds to English copular constructions, while the English form that corresponds to *tomber bien* is another verb (*come*), but with the same kind of general meaning. Both French *tomber* and English *come* are unaccusative verbs of directed motion.

Finally, looking into the translations of non-matching predicates in the parallel corpus, we notice that the English translations of French expressions are not always consistent with what has been identified as an equivalent English expression, as well as between themselves. This is especially the case with the group 4 mismatches involving impersonal intransitive verbs (*il s'agit de*, *il se peut que*, *il convient*). On the other hand, the transitive predicates of the group 4 are translated with more consistency. A detailed list of all the predicates found in the corpus and their translations can be found in the Appendix.

## 5. UNIFYING REPRESENTATION OF THE NON-MATCHING PREDICATES

As we show in the previous section, the structures associated with French predicates identified as not matching with English predicates during the annotation still retain certain level of parallelism with the corresponding English structures. In most of the cases, it is precisely the number and the meaning of the predicates' arguments that remain unchanged. Variation can be identified at other levels of the structure. It is limited to two domains: the choice of the lexemes which encode the impoverished meaning of the predicates and the choice of lexical category of the predicating lexical unit.

For cross-linguistic mapping of the predicate-argument structures which involve variation in one of the two domains, the representation of the structure needs to be slightly more abstract than the one that is currently used in PropBank. In this section, we propose the representations of the structures which are valid cross-linguistically introducing only minimal generalisations needed to address the observed variation.



### 5.1. Parallel structures

In the case where the heading verbs are not translations of each other, while the rest of the structure is preserved across languages (group 1, described in 4.2.1), we propose assigning a special label to the verbs. Since the meaning of these verbs is impoverished in these usages, no specific verb sense label can account for it. The label on these verbs needs to express their general meaning and it needs to be applicable to multiple verbs. Such a representation of the sentence in (4) is given in (15).

- (15) a. [ARG0 Nous] [REL-CAUS tiron] [ARG1 les **leçons** du passé]. (French)  
 b. [ARG0 We] [REL-CAUS learn] [ARG1 the **lessons** of the past]. (English)

The semantically impoverished verb in French is assigned a more abstract label CAUS. This label marks two characteristics of the predicate at the same time: the fact that it is not an ordinary predicate and its general meaning. All the impoverished verbs in our sample have the same general meaning – vaguely causative – but other labels could be used for other general meanings that can potentially occur.

This approach would require a set of abstract labels to be defined within the framework of PropBank in addition to the verb sense labels. These labels would resemble in some aspect the FrameNet labels, since they would express an abstract layer of the meaning of the verbs and since they would be applied to multiple lexical units. The difference is that these labels would encode a more abstract meaning than is the case with the FrameNet labels. Also, they would be used only for a very limited set of lexical items in special contexts where their meaning is impoverished.

A special case of almost parallel structures is represented in (8), where the French pronominal verb corresponds to the English transitive verb. A parallel representation of this sentence, given in (16), requires treating the pronominal clitic in French (*m* in (16a)) as one of the verb's arguments and assigning it the same label as to the corresponding argument in English (ARG1 in (16b)).

- (16) a. [ARG0 Je] [ARG1 **m'**] [REL-EXPRESS.01 **exprime**] [ARG2 sur le texte de M. L]. (French)  
 b. [ARG0 I] [REL-EXPRESS.01 **express**] [ARG1 **my opinion**] [ARG2 on the Mr L.'s document]. (English)

The meaning of the ARG1 in the English sentence can be interpreted as non-specific or general, but this has no consequences for the annotation framework, since the argument labels are already general.

### 5.2. Category changing structures

The non-parallel structures that involve a change in the lexical category of the predicating word require a stronger deviation from the standard PropBank representation. Most of these cases involve a predicate that is expressed synthetically in one language (as a single verb), while it is expressed analytically in the other language (as a combination of a verb and a predicating complement). The corresponding lexical units in these cases are the synthetic verb in one language and the verb's complement in the analytical expression in the other language.

A possible way of representing the cross-linguistic parallelism in the predicate-argument structure of these expressions is to label the corresponding lexical items with a predicate label, leaving the heading verb of the analytical expression without a label, as illustrated in (17).

- (17) a. [ARG0 Le marché] **a** [REL-NEED.01 **besoin**] [ARG1 de règles]. (French)  
 b. [ARG0 The market] [REL-NEED.01 **needs**] [ARG1 rules]. (English)

The verbal head of the French analytical predicate (a) is not assigned a predicate label. Instead, the predicate is the noun heading its complement (*besoin*), which is, at the same time, lexical counterpart of the verb in English (*need*). The unlabeled verb in this case would be treated as a functional word which has no semantic arguments - a lexicalized light verb (Hale & Keyser 1993, Grimshaw & Mester 1988, Kearns 2002).

Such representation is not possible in the current PropBank setting due to two limitations. First, all the verbs except the copular *be* are considered as predicates. Leaving a non-copular verb without a label would go against this principle. Moreover, a strictly functional treatment of light verbs is the subject to debate in the literature, with some authors arguing in favour of such approach for certain verbs (Wierzbicka 1982, Kearns 2002, Grimshaw & Mester 1988) and others providing evidence of semantic content of light verbs (Brugman 2001). Second, only verbs are considered as predicates in PropBank, which is why the frames are not specified for other lexical categories. This means that an existing verbal frame in the English resource would need to be adapted to be used with a French nominal predicate. The limitation is even more important in predicates which are expressed in an analytical form in English, such as *be enough* in (18b), corresponding to the French *suffit* in (18a). The argument structure of the English adverb *enough* is not specified at all in PropBank's frame files.

- (18) a. Il [REL-ENOUGH **suffit**] [ARG0 de lire le programme]. (French)  
 b. It **is** [REL-ENOUGH **enough**] [ARG0 to read the manifesto]. (English)

Both limitations affect potentially parallel representation of the predicates which are expressed analytically in both languages, but for which the English expression includes the copular *be*, such as *avoir qqch. comme objectif* in (7) or *avoir retard, prendre conscience de* in (11). As an illustration for this case, we propose a representation (19) of the sentence in (6).

- (19) a. [ARG0 L' **Union**] ne peut pas **avoir** comme [REL-OBJECTIVE **objectif** principal] [ARG1 de **réduire** le niveau global des aides]. (French)  
 b. [REL-OBJECTIVE The main **objective**] [ARG0 of the **Union**] cannot **be** [ARG1 to **reduce** the overall level of aid]. (English)

The French verb *avoir* is not assigned a predicate label in this example. Instead, the label is assigned to the head of its predicative complement (*objectif*), which corresponds to the head of the subject phrase in the English sentence (the noun *objective*). A parallel representation of these structures requires not just treating the French verb as a functional word, but also specifying the predicate label of a nominal predicate.

### 5.3. Expressions with no direct translations

Finally, the cases identified as the strongest mismatches in our classification can be resolved by applying one of the described approaches too, as illustrated in (20). The only difference is that the English label will not be a direct translation of the French word in question. This, however, should not pose a problem for the result of annotation. The label does not have to be a direct translation. It only needs to be consistent (and consistently found in a corpus) with the word.

- (20) a. [ARG0 Les dirigeants politiques] doivent **faire** [REL-SHOW.01 **preuve**] [ARG1 d'un réel courage]. (French)  
 b. [ARG0 Political leaders] must [REL-SHOW.01 **show**] [ARG1 real courage]. (English)

### 5.4. Concluding remarks

The analysis of the mismatching expressions points to only two sources of the observed variation.

In the case of mismatching expressions with lexical categories preserved across languages (groups 1 and 2), it is the impoverished meaning of the verb that increases the potential variation in the choice of lexical item. Using a special, abstract, label for verbs with impoverished meaning would ensure a more adequate and cross-linguistically valid annotation framework.

The cases of category changing predicates show that the possibility of expressing the same predicate with different lexical categories (verb, noun, or adjective) increases the chances of different realizations across languages. Specifying predicate-argument structures for all predicating lexical items, including verbs, nouns, adjectives, and adverbs is important for making an annotation framework directly portable across languages.

The divergent cases identified in our study were not due to a truly different predicate-argument structure, but rather to the limitations imposed by the annotation framework.

## 6. CONCLUSIONS

A translation is an implicit representation of the meaning of a sentence. Forcing annotations, which are explicit representations, to be similar across languages by studying divergences and modifying annotations to address these divergences is a way of making explicit the components of meaning that are expressed in a predicate.

The general linguistic conclusion of our investigation is that the cross-linguistic variation in the predicate-argument structure concerns the lexical realisations of the predicates and not the argument structure itself. Even for the cases identified as mismatches in our study, the number and the kind of arguments are unchanged across languages. The mismatches are due to the different choices the two languages make with respect to lexical items that realise the meaning of semantically impoverished verbs, and to the lexical category of the predicating word. Our analysis also suggests that the structural correspondence fails more when a clause contains a verb with less specific meaning. This correlation between specificity of meaning and structural realisation is a confirmation of the cross-linguistic validity of those approaches, such as (Levin & Rappaport Hovav 2005), that hypothesize a direct correspondence between the components of meaning of a predicate and the syntactic realisation of the arguments.

Applying a specific annotation framework such as PropBank in a cross-linguistic study proved plausible. A great majority of French predicates found in our corpus (95%) corresponded directly to an English verb sense specified in PropBank. Our findings about the sources of the mismatches in the remaining predicates indicate that using a small set of special labels for verbs with impoverished meaning and providing annotation for all categories of predicating words (verbs, nouns, and adjectives) would improve the empirical adequacy and cross-linguistic validity of this framework.

## REFERENCES

- Babko-Malaya, O. (2005) PropBank annotation guidelines.
- Baker, C.F., C. J. Fillmore & J. B. Lowe (1998) "The Berkeley FrameNet project", in *Proceedings of the 36<sup>th</sup> Annual Meeting of the Association for Computational Linguistics and 17<sup>th</sup> International Conference on Computational Linguistics*, ACL/Morgan Kaufmann Publishers, Montreal, Canada, 86-90.
- Brugman, C. (2001) "Light verbs and polysemy", *Language Science* 23, 551-578.
- Burchardt, A., K. Erk, A. Frank, A. Kowalski, S. Pado & M. Pinkal (2009) "FrameNet for the semantic analysis of German: Annotation, representation and automation", in Hans Boas, editor, *Multilingual FrameNets in Computational Lexicography: Methods and applications*, Mouton de Gruyter, 209-244.
- Burnard, L. (2007) "Reference guide for the British National Corpus (XML edition)", URL <http://www.natcorp.ox.ac.uk/XMLedition/URG/>.
- Candito, M. H. & B. Crabbé & P. Denis and F. Guérin (2009) "Analyse syntaxique du Français: des constituants aux dépendances", in *Actes de la Conférence sur le Traitement Automatique des Langues Naturelles (TALN'09)*, Senlis, France.
- Dowty, D. (1991) "Thematic proto-roles and argument selection", *Language* 67(3), 547-619.
- Fellbaum, C. (1998) *WordNet, an electronic lexical database*, MIT Press.
- Fillmore, C. J. (1968) "The case for case", in Emmon Bach and Robert T. Harms, editors, *Universals in linguistic theory*, Holt, Rinehart and Winston, New York, 1-88.
- Fillmore, C. J. (1982) "Frame semantics", in *Linguistics in the Morning Calm*, Hanshin Publishing Co., Seoul, 111-137.
- Fung, P., Z. Wu, Y. Yang & D. Wu (2007) "Learning bilingual semantic frames: Shallow semantic parsing vs. semantic role projection", in *11<sup>th</sup> Conference on Theoretical and Methodological Issues in Machine Translation (TMI 2007)*, Skovde, Sweden, 75-84.
- Grimshaw, J. & A. Mester (1988) "Light verbs and theta-marking", *Linguistic Inquiry* 19, 205-232.
- Hale, K. & S. J. Keyser (1993) "On argument structure and the lexical representation of syntactic relations", in Kenneth Hale and Samuel Jay Keyser, editors, *The View from Building 20*, MIT Press, 53-110.
- Jackendoff, R. (1990) *Semantic Structures*, MIT Press, Cambridge, Mass.
- Kearns, K. (2002) *Light verbs in English*, manuscript.
- Kayne, R. (2005) "Some notes on comparative syntax, with special reference to English and French", in *The Oxford Handbook of Comparative Syntax*, Oxford University Press.
- Kipper, K. S. (2005) *VerbNet: A broad-coverage, comprehensive verb lexicon*, PhD thesis, University of Pennsylvania.
- Koehn, P. (2005) Europarl: A parallel corpus for statistical machine translation, in *Proceedings of MT Summi*, Phuket, Thailand.
- Levin, B. (1993) *English verb classes and alternations: a preliminary investigation*, The University of Chicago Press, Chicago.

- Levin, B. & M. Rappaport Hovav (2005) *Argument realization*, Cambridge University Press, Cambridge.
- Loper, E., S. T. Yi & M. Palmer (2007) "Combining lexical resources: Mapping between PropBank and VerbNet", in *Proceedings of the 7th International Workshop on Computational Linguistics*, Tilburg, the Netherlands.
- Pado, S. (2007) *Cross-Lingual Annotation Projection Models for Role-Semantic Information*. PhD thesis, Saarland University.
- Palmer, M., D. Gildea & P. Kingsbury. (2005) "The Proposition Bank: An annotated corpus of semantic roles", *Computational Linguistics* 31(1), 71-105.
- Pajas, P. & J. Stepanek (2008) "Recent advances in a feature-rich framework for treebank annotation", in *Proceedings of the 22<sup>nd</sup> International Conference on Computational Linguistics (Coling 2008)*, Manchester, UK, 673-680.
- Reinhart, T. (2002) "The theta system – An overview", *Theoretical linguistics* 28, 229-290.
- Titov, I. & J. Henderson (2007) "A Latent Variable Model for Generative Dependency Parsing", in *Proceedings of the International Conference on Parsing Technologies (IWPT-07)*, Prague, Czech Republic, 44-155.
- Wierzbicka, A. (1982) "Why can you have a drink when you can't \*have an eat?", *Language* 58 (4), 753-799.

## APPENDIX

## Transitive

French expression	English counterpart	French instance	English translation
<b>Preserved lexical and syntactic categories (Group 1)</b>			
faire la lumière	shed light	faire toute la lumière	shed some light
faire alliance	form an alliance	un gouvernement fait alliance avec des néofascistes	a government forms an alliance with neo-Fascists
faire pression	put pressure	un moyen de faire pression sur ces pays	some way of putting pressure on such countries
faire appel	make an appeal	que les investissements... puissent faire appel à l' IFOP	that appeals can be made to FIFG for investments...
attirer l'attention	draw attention	je voudrais attirer l' attention sur certains aspects des résolutions	I should like to draw attention to some of the aspects of the resolutions
tirer la leçon	learn the lesson	faisons en sorte... de tirer cette fois réellement la leçon	Let us ensure...that this time we really do learn the lesson
		tirons les leçons du passé	Let us learn the lessons of the past
		une leçon fondamentale que je voudrais que nous tirions	one basic lesson I would like us to learn
céder la place	make room	pour céder la place à un usage plus vaste du droit communautaire	permitting greater use of Community law
avoir lieu	take place	lorsque le tremblement de terre a eu lieu en Grèce	When we had the earthquake in Greece
		sans qu' un débat ou un vote n' ait eu lieu	without being discussed or voted on
avoir rien à voir	have nothing to do	la plainte actuelle n' a rien à voir en ampleur avec la plainte originelle	this complaint is insignificant compared to the original complaint
<b>Preserved lexical, changed syntactic categories (Group 2)</b>			
avoir qqch comme objectif	be the objective	l' Union ne peut pas avoir comme objectif principal de réduire le niveau global des	The main objective of the Union ... cannot be to reduce the overall level of
		cette formation...devait avoir pour objectif la connaissance des dispositions législatives	this professional training would be targeted at awareness of the terms of the legislation
faire l'objet de	be the objective	quatre...pays vont faire l' objet d' une décision	the Commission is to decide on the matter for four more countries

donner à qqn. à penser	make sbd. think	Cela me donne à penser que ...	This suggests to me that...
laisser qqn. sceptique	be sceptical	la réponse...me laisse un peu sceptique	I am a little sceptical of the reply
porter un jugement	make an assessment	le jugement que vous portez...sur ce sixième rapport	the assessment you have made of this sixth...report
rediscuter	discuss again	Nous discuterons à l'avenir d'une énième catastrophe maritime.	We will be back in the future discussing yet another maritime disaster.
tenir compte de	take into account	cette entité...tient compte des réalités hydrologiques	a body that ... takes account of the real water situation
		une histoire propre dont il nous faut tenir compte	our histories to live with
		il est préférable de tenir compte des invitations	better to heed calls
		compte tenu de l'importance de leur impact	because they are important
		vous alliez en tenir compte	you will act on them
		nous devrions en tenir compte	we should get it into perspective
<b>Lexical categories changed (Group 3)</b>			
prendre conscience de	be/become aware	j' ai pris... conscience de la détermination	I have become...aware of the determination
avoir besoin de	need	Le marché ...a besoin de Règles.	the market needs rules
		la Commission a... besoin de la possibilité de négocier	the Commission...need the scope to negotiate
		Ce dont nous avons besoin maintenant , c' est d' agir	What is now needed is action
		Nous avons besoin d' un organe responsable de la législation alimentaire	We need a body of food law
		nous avons besoin d' une direction politique	we must find political leadership
		nous avons besoin d' une Commission dotée de	we need a properly resourced Commission
		les moyens dont elle a besoin	the resources that the Commission needs
		nous avons besoin d' une bonne dose de pragmatisme	What is needed ... is a healthy dose of pragmatism

rendre compte	be accountable	leur obligation de rendre compte...des aides	their obligation to be accountable...for the aid
		es responsables doivent rendre des comptes	those responsible are liable
porter remède	remedy	Il faut...porter remède à cette carence	this omission must be remedied
donner suite à	follow up	Il est important que la Commission donne suite à ces recommandations	and it is important that it [response] is followed through
		vous n' y avez pas donné suite	they have not been acted upon
		Nous voulons qu' à l' avenir, on donne bien plus cette suite à nos recommandations.	We want to see that follow-through to a much greater extent in future.
y faire face	face	un instrument pour y faire	an instrument to deal with it
		pour y faire face...	...to deal with such cases
faire état	state	un registre qui fasse état du montant des aides	a register specifying the amount of state aid
faire défaut	lack	des normes européennes font défaut	there are no applicable European standards met
mettre qqch en exergue	highlight	de mettre en exergue les conséquences	to highlight the result
avoir trait de	have to do	les craintes concernant les armes à uranium appauvri ont trait à la manière dont celles-ci sont utilisées	the concern with depleted uranium weapons is about the way in which they are used
		...tous deux ayant trait au vote	These are both about vote.
avoir retard	be late	nous avons du retard	We are coming from behind

#### No direct translation (Group 4)

faire preuve de	show	nous devons aussi faire preuve de responsabilité	we also have a responsibility here
		Les dirigeants politiques doivent faire preuve d' un réel courage	Political leaders must show real courage
		vous avez fait preuve de conviction	you have shown conviction
		la Commission et son président ont fait preuve d' une retenue louable	the Commission and its President exercised commendable self-restraint



		les avocats-conseils... peuvent faire preuve d ' une telle obstination	the legal advisers...could demonstrate such stubbornness
prendre la parole	speak	la Présidente...m' a demandé de prendre la parole	the invitation...by the President...to speak
prendre la peine	bother	Je n' ai pas l' habitude de prendre la peine de féliciter les gens	I do not usually bother congratulating people
porter atteinte	undermine	il pourrait...porter atteinte au super-État uniformisateur	it might...[be] attacking the standardising superstate
mettre qqch en oeuvre	implement	les réformes...soient pleinement mises en uvre	the reforms...are fully implemented
se faire un plaisir	be happy	je me ferai un plaisir de vous exposer nos vues	I will be happy to present our viewpoint to you
remettre qqch en état	repair	de la remettre en état le plus rapidement possible	to repair them as quickly as possible
remettre qqch. en cause	call stg. into question	rien ne justifie qu' elle soit remise en cause	there is nothing to justify any reconsideration thereof
assigner qqn. à résidence	put sbd. under house arrest	M. G. est...assigné à résidence	He is...in his house

## Intransitive

French expression	English counterpart	French instance	English translation
<b>Preserved lexical, changed syntactic categories (Group 2)</b>			
s'exprimer	express an opinion	je m' exprime ce matin, plus particulièrement, sur le texte de M. L.	I am speaking this morning with particular reference to Mr L.'s document
<b>Lexical categories changed (Group 3)</b>			
se mettre d'accord	agree	nous nous sommes mis d' accord hier pour inscrire le rapport...à l' ordre du jour d' aujourd' hui	we agreed yesterday to have the Bourlanges report on today 's agenda
se réjouir	be happy	Les citoyens...vont s' en réjouir	citizens...will be happy about [it]
il suffit	be enough	il suffit de lire le programme	it is enough to read the manifesto
<b>No direct translation (Group 4)</b>			
se prononcer	give an opinion	Je ne possède donc pas les compétences requises	I am therefore not qualified to give an opinion

		pour me prononcer	
tenir de	be	tout cela tient de la rhétorique au vu des obstacles	all this...is so much rhetoric when we look at the...obstacles
se féliciter	be pleased	je me félicite...du fait que la présidence portugaise fasse de la justice	I...welcome the fact that the Portuguese Presidency is making justice
se rendre à	go to	...en me rendant moi-même devant chacun des parlements nationaux	by visiting each of the national parliaments
tomber bien	come in a good moment	notre débat tombe bien ce soir	our debate has come at an opportune time
il s'agit de	be about	il s'agit là de questions... importantes	they are...important issues
		il s'agit des PME	the issue of SMEs
		il s'agit d'un état de fait très insatisfaisant	a very unsatisfactory state of affairs
		lorsqu'il s'agit de sécurité maritime	when it comes to maritime safety
		lorsqu'il s'agit de la Commission	when it is put to the Commission
		Il s'agit bien d'une catastrophe écologique	What we have here is an ecological disaster
		Il s'agit là d'une grande exagération	That is a gross exaggeration
		lorsqu'il s'agit de la façon dont nous traitons les réfugiés	when it comes to our treatment of refugees
		Il s'agit simplement pour les États de jouer les vases communicants	It is nothing more than the States playing
		Il s'agit de ce phénomène de solidarité qui est à l'origine même du modèle social européen.	It is precisely this phenomenon of solidarity which lies behind the European social model.
il convient	should	Il convient de s'occuper de la menace	The issue of the threat... needs to be addressed
		Il convient de reconnaître que...	It is fair to say...

		Il convient de s'occuper de ce problème	This needs to be looked into
		il convient de prendre des mesures	something needs to be done
il se peut que	it is possible	il se peut qu' elle ait ruiné de nombreuses existences	it may have ruined many human livelihoods