# Cross-lingual Validity of PropBank in the Manual Annotation of French

**Lonneke van der Plas** **Tanja Samardžić** **Paola Merlo**

Linguistics Department
University of Geneva
Rue de Candolle 5, 1204 Geneva
Switzerland
{Lonneke.vanderPlas,Tanja.Samardzic,Paola.Merlo}@unige.ch

## Abstract

Methods that re-use existing mono-lingual semantic annotation resources to annotate a new language rely on the hypothesis that the semantic annotation scheme used is cross-lingually valid. We test this hypothesis in an annotation agreement study. We show that the annotation scheme can be applied cross-lingually.

## 1 Introduction

It is hardly a controversial statement that elegant language subtleties and powerful linguistic imagery found in literary writing are lost in translation. Yet, translation preserves enough meaning across language pairs to be useful in many applications and for many text genres.

The belief that this layer of meaning which is preserved across languages can be formally represented and automatically calculated underlies methods that use parallel corpora for the automatic generation of semantic annotations through cross-lingual transfer (Padó, 2007; Basili et al., 2009).

A methodology similar in spirit — re-use of the existing resources in a different language — has also been applied in developing manually annotated resources. Monachesi et al. (2007) annotate Dutch sentences using the PropBank annotation scheme (Palmer et al., 2005), while Burchardt et al. (2009) use the FrameNet framework (Fillmore et al., 2003) to annotate a German corpus. Instead of building special lexicons containing the specific semantic information needed for the annotation for each language separately, which is a complex and time-consuming endeavour in itself, these approaches rely on the lexicons already developed for English.

In this paper, we hypothesize that the level of abstraction that is necessary to develop a semantic lexicon/ontology for a *single language* based on *observable linguistic behaviour* — that is a mono-lingual, item-specific annotation — is cross-linguistically valid. We test this hypothesis by manually annotating French sentences using the PropBank frame files developed for English.

It has been claimed that semantic parallelism across languages is smaller when using the PropBank semantic annotations instead of the FrameNet scheme, because FrameNet is more abstract and less verb-specific (Padó, 2007). We are working with the PropBank annotation scheme, contrary to other works that use the FrameNet scheme, such as Padó (2007) and Basili et al. (2009). We choose this annotation for two main reasons. First, the primary use of our annotation is to serve as a gold standard in the task of syntactic-semantic parsing. FrameNet does not have a properly sampled hand-annotated corpus of English, by design. So we cannot use it for this task. Second, in Merlo and Van der Plas (2009), the semantic annotations schemes of PropBank and VerbNet (Kipper, 2005) are compared, based on annotation of the SemLink project (Loper et al., 2007). The authors conclude that PropBank is the preferred annotation for a joint syntactic-semantic setting.

If the PropBank annotation scheme is cross-lingually valid, annotators can reach a consensus and can do so swiftly. Thus, cross-lingual validity is measured by how well-defined the manual annotation task is (inter-annotator agreement) and by how hard it is to reach an agreement (pre- and post-consensus inter-annotator agreement). In addition, we measure the impact of the level of abstraction of the predicate labels. Conversely, how often labels do not transfer and distributions of disagreements are indicators of lack of parallelism across languages that we study both by quantitative and qualitative analysis.

To preview the results, we find that the PropBank annotation scheme developed for English can be applied for a large portion of French sen-

*Proceedings of the Fourth Linguistic Annotation Workshop, ACL 2010*, pages 113–117,
Uppsala, Sweden, 15-16 July 2010. ©2010 Association for Computational Linguistics

tences without adjustments, which confirms its cross-lingual validity. A high level of inter-annotator agreement is reached when the verb-specific PropBank labels are replaced by less fine-grained verb classes after annotating. Non-parallel cases are mostly due to idioms and collocations.

## 2 Materials and Methods

Our choices of formal representation and of labelling scheme are driven by the goal of producing useful annotations for syntactic-semantic parsing in a setting based on an aligned corpus. In the following subsections we describe the annotation scheme and procedure, the corpus, and phases of annotation.

### 2.1 The PropBank Annotation Framework

We use the PropBank scheme for the manual annotations. PropBank is a linguistic resource that contains information on the semantic structure of sentences. It consists of a one-million-word corpus of naturally occurring sentences annotated with semantic structures and a lexicon (the PropBank frame files) that lists all the predicates (verbs) that can be found in the annotated sentences and the sets of semantic roles they introduce.

Predicates are marked with labels that specify the sense of the verb in the particular sentence. Arguments are marked with the labels A0 to A5. The labels A0 and A1 have approximately the same value with all verbs. They are used to mark instances of typical AGENTS (A0) and PATIENTS (A1). The value of other numbers varies across verbs. Modifiers are annotated in PropBank with the label AM. This label can have different extensions depending on the semantic type of the constituent, for example *locatives* and *adverbials*.

### 2.2 Annotation Procedure

Annotators have access to PropBank frame files and guidelines adapted for the current task. The frame files provide verb-specific descriptions of all possible semantic roles and illustrate these roles with examples as shown for the verb *paid* in (1) and the verb senses of *pay* in Table 1. Annotators need to look up each verb in the frame files to be able to label it with the right verb sense and to be able to allocate the arguments consistently.

(1) [$_{A0}$ The Latin American nation] has [$_{REL-PAY.01}$ paid] [$_{A1}$ very little] [$_{A3}$ on its debt] [$_{AM-TMP}$ since early last year].

| Frame | Semantic roles |
|---|---|
| pay.01 | A0: payer or buyer<br>A1: money or attention<br>A2: person being paid, destination of attention<br>A3: commodity, paid for what |
| pay.02<br>*pay off* | A0: payer<br>A1: debt<br>A2: owed to whom, person paid |
| pay.03<br>*pay out* | A0: payer or buyer<br>A1: money or attention<br>A2: person being paid, destination of attention<br>A3: commodity, paid for what |
| pay.04 | A1: thing succeeding or working out |
| pay.05<br>*pay off* | A1: thing succeeding or working out |
| pay.06<br>*pay down* | A0: payer<br>A1: debt |

Table 1: The PropBank lexicon entry for *pay*.

In our cross-lingual setting, annotators used the English PropBank frame files to annotate the French sentences. This means that for every predicate they find in the French sentence, they need to translate it, and find an English verb sense that is applicable to the French verb. If an appropriate entry cannot be found in the frame files for a given predicate, the annotator is instructed to use the "dummy" label for the predicate and fill in the roles according to their own insights.

For the annotation of sentences we use an adaptation of the user-friendly, freely available Tree Editor (TrEd, Pajas and Štěpánek, 2008). 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.

The decision to show syntactic information is merely driven by the fact that we want to guide the annotator in selecting the heads of phrases during the annotation process. The sentences are parsed by a syntactic parser (Titov and Henderson, 2007) that we trained on syntactic dependency annotations for French (Candito et al., 2009). Although the parser is state-of-the-art (87.2% Labelled Attachment Score), in case of parse errors, we ask annotators to ignore the errors of the parser and put the label on the actual head.

### 2.3 Corpus

We selected the French sentences for the manual annotation from the parallel Europarl corpus (Koehn, 2005). Because translation shifts are known to pose problems for the automatic cross-lingual transfer of semantic roles (Padó, 2007) and for machine translation (Ozdowska and Way, 2009), and these are more likely to appear in indirect translations, we decided to select only those parallel sentences, for which we can infer from the labels used in Europarl that they are direct translations from English to French, or vice versa. We selected 1040 sentences for annotation (40 in total for the two training phases, 100 for calibration, and 900 for the main annotation phase.)[1]

### 2.4 Annotation Phases

The training procedure described in Figure 1 is inspired by the methodology indicated in Padó (2007). A set of 130 sentences were annotated manually by four annotators with very good proficiency in both French and English for the training and the calibration phase. The remaining 900 sentences are annotated by one annotator (out of those four), a trained linguist. Inter-annotator agreement was measured at several points in the annotation process marked with an arrow in Figure 1. The guidelines were adjusted after the training phase.

- Training phase
  -TrainingA: 10 sentences, all annotators together
  -TrainingB: 30 sentences, all annotators individually⇐
  -Reach consensus on Training B ⇐
- Calibration phase
  -100 sentences by main annotator, one third of those by each of the other 3 annotators ⇐
- Main annotation phase
  -900 sentences by main annotator

Figure 1: The annotation phases.

## 3 Results

Cross-lingual validity is measured by comparing inter-annotator agreement at several stages in the annotation, by measuring the agreement on less specific predicate labelling, and by a quantitative and qualitative analysis of non-parallel cases.

### 3.1 Inter-annotator Agreement for Several Annotation Phases

To assess the quality of the manual annotations we measured the agreement between annotators as the average F-measure of all pairs of annotators after each phase of the annotation procedure.[2] The first row of Table 2 shows that the task is hard. But the difference between the first row and the second row shows that there were many differences between annotators that could be resolved. After discussions and individual corrections the scores are between 91% and 95%. This indicates that the task is well-defined. Row three shows that the agreement in the calibration phase increases a lot compared to the last training phase (row 1). This might in part be due to the fact that the guidelines were adjusted by the end of the training phase, but could also be because the annotators are getting more acquainted to the task and the software.

|  | Predicates |  | Arguments |  |
|---|---|---|---|---|
|  | Lab. F | Unl. F | Lab. F | Unl. F |
| TrainingB | 46 | 85 | 62 | 75 |
| TrainingB(cons.) | 95 | 97 | 91 | 95 |
| Calibration | 59 | 93 | 69 | 84 |

Table 2: Percent inter-annotator agreement (F-measure) for labelled/unlabelled predicates and for labelled/unlabelled arguments

As expected, because annotators used the English PropBank frame files to annotate French verbs, the task of labelling predicates proved more difficult than labelling semantic roles. It results in the lowest agreement scores overall. In the following subsections we study the sources of disagreement in predicate labelling in more detail.

### 3.2 Inter-annotator Agreement in Predicate Labellings

Predicate labels in PropBank apply to particular verb senses, for example *walk.01* for the first sense of the verb *walk*. Even though the senses are coarser than, for example, the senses in WordNet (Fellbaum, 1998), the labels are rather specific. This specificity possibly poses problems when working in a cross-lingual setting.

We compare the agreement reached using PropBank verb sense labels with the agreement reached using the verb classifications from VerbNet (Kipper, 2005) and the mapping to PropBank labels as provided in the type mappings of the SemLink project[3] (Loper et al., 2007). If two annotators used two different predicate labels to annotate the

[1] As usual practice in preprocessing for automatic alignment, the datasets were tokenised and lowercased and only sentence pairs corresponding to a 1-to-1 alignment with lengths ranging from 1 to 40 tokens on both French and English sides were considered.

[2] It is a known fact that measuring annotator agreement using the kappa score is problematic in categorisation tasks that do not have a fixed number of items and categories (Burchardt et al., 2006). The F-measure is a well-known measure used for the evaluation of many task such as syntactic-semantic parsing, the task that is the motivation for this paper. The choice of the F-measure makes the comparison to the performance of the future parser easier.

[3] (http://verbs.colorado.edu/semlink/)

same verb, but those verb senses belong to the same verb class, we count those as correct[4].

The average inter-annotator agreement is relatively low when we compare the annotations on the PropBank verb sense level: 59%. However, at the level of verb classes, the inter-annotator agreement increases to 81%. This raises the issue of whether we should not label the predicates with verb classes instead of verb senses. By using PropBank labels for the manual annotation and replacing these with verb classes in post-processing, the benefits are two-fold: We are able to reach a high level of cross-lingual parallelism on the annotations, while keeping the manual annotation task as specific and less abstract as possible.

### 3.3 Analysis of Non-Parallel Cases

For a single annotator, the main measure of cross-lingual validity is the percentage of dummy predicates in the annotation. In the sentences from the calibration and the main annotation phase from the main annotator (1000 sentences in total), we find 130 predicates (tokens) for which the annotator used the "dummy" label.

Manual inspection reveals that the "dummy" label is mainly used for French multi-word expressions (82%), most of which can be translated by a single English verb (47%), whereas others cannot, because they are translated by a combination that includes a form of 'be' that is not annotated in PropBank (25%). The 47% of multi-word expressions that receive the "dummy" label show the annotator's reluctance to put a single verb label on a French multi-word expression. The annotation guidelines could be adapted to instruct annotators not to hesitate in such cases.

Similarly, collocations and idiomatic expressions are the main sources of disagreement in predicate labellings among annotators. We can conclude that, as shown in studies on other language pairs (Burchardt et al., 2009), collocations and idiomatic expressions were identified as verb uses where the verb's predicate label cannot be transferred directly from one language to another.

## 4 Discussion and Related Work

Burchardt et al. (2009) use English FrameNet to annotate a corpus of German sentences manually. They find that the vast majority of frames can be applied to German directly. However, around one third of the verb senses identified in the German corpus were not covered by FrameNet. Also, a number of German verbs were found to be underspecified. Finally, some problems related to treating particular verb uses were identified, such as idioms, metaphors, and support verb constructions.

Monachesi et al. (2007) use PropBank labels for semi-automatic annotation of a corpus of Dutch sentences. Semantic roles were first annotated using a rule-based semantic parser and then corrected by one annotator. Although not all Dutch verbs could be translated to an equivalent verb sense in English, these cases were assessed as relatively rare. What proved to be problematic was identifying the correct label for modifiers.

Bittar (2009) makes use of cross-lingual lexical transfer in annotating French verbs with event types, by adapting a small-scale English verb lexicon with specified event structure (TimeML).

The inter-annotator agreement in labelling predicates reported in Burchardt et al. (2009) reaches 85%, while our best score (when falling back to verb classes) is 81%. However, unlike Burchardt et al. (2009) we did not introduce any new French labels. We find, like Monachesi et al. (2007), that non-parallel cases are less frequent than what is reported in Burchardt et al. (2009), which could be due to the properties of the annotations schemes.

## 5 Conclusions

We can conclude that the general task of annotating French sentences using English PropBank frame files is well-defined. Nevertheless, it is a hard task that requires linguistic training. With respect to the disagreements on labelling predicates, we can conclude that a large part can be resolved if we compare the annotations at the level of verb classes instead of at the very fine-grained level of verb senses. Non-parallel cases are mostly due to idioms and collocations. Their rate is relatively low and can be further reduced by adapting annotation guidelines.

## Acknowledgments

The research leading to these results has received funding from the EU FP7 programme (FP7/2007-2013) under grant agreement nr 216594 (CLASSIC project: www.classic-project.org). We would like to thank Goljihan Kashaeva and James Henderson for valuable comments.

[4] 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.

## References

R. Basili, D. De Cao, D. Croce, B. Coppola, and A. Moschitti, 2009. *Computational Linguistics and Intelligent Text Processing*, chapter Cross-Language Frame Semantics Transfer in Bilingual Corpora, pages 332–345. Springer Berlin / Heidelberg.

A. Bittar. 2009. Annotation of events and temporal expressions in French texts. In *Proceedings of the third Linguistic Annotation Workshop (LAW III)*, pages 48–51, Suntec, Singapore.

A. Burchardt, K. Erk, A. Frank, A. Kowalski, S. Padó, and M. Pinkal. 2006. The SALSA corpus: a German corpus resource for lexical semantics. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC 2006)*, pages 969–974, Genoa, Italy.

A. Burchardt, K. Erk, A. Frank, A. Kowalski, S. Pado, and M. Pinkal, 2009. *Multilingual FrameNets in Computational Lexicography: Methods and Applications*, chapter FrameNet for the semantic analysis of German: Annotation, representation and automation, pages 209–244. De Gruyter Mouton, Berlin.

M.-H. Candito, B. Crabbé, P. Denis, and F. Guérin. 2009. Analyse syntaxique du français : des constituants aux dépendances. In *Proceedings of la Conférence sur le Traitement Automatique des Langues Naturelles (TALN'09)*, Senlis, France.

C. Fellbaum. 1998. WordNet, an electronic lexical database. MIT Press.

C. J. Fillmore, R. Johnson, and M.R.L. Petruck. 2003. Background to FrameNet. *International journal of lexicography*, 16.3:235–250.

K. Kipper. 2005. *VerbNet: A broad-coverage, comprehensive verb lexicon*. Ph.D. thesis, University of Pennsylvnia.

P. Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of the MT Summit*, pages 79–86, Phuket, Thailand.

E. Loper, S-T Yi, and M. Palmer. 2007. Combining lexical resources: Mapping between PropBank and VerbNet. In *Proceedings of the 7th International Workshop on Computational Semantics (IWCS-7)*, pages 118–129, Tilburg, The Netherlands.

P. Merlo and L. van der Plas. 2009. Abstraction and generalisation in semantic role labels: PropBank, VerbNet or both? In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 288–296, Suntec, Singapore.

P. Monachesi, G. Stevens, and J. Trapman. 2007. Adding semantic role annotation to a corpus of written Dutch. In *Proceedings of the Linguistic Annotation Workshop (LAW)*, pages 77–84, Prague, Czech republic.

S. Ozdowska and A. Way. 2009. Optimal bilingual data for French-English PB-SMT. In *Proceedings of the 13th Annual Conference of the European Association for Machine Translation (EAMT'09)*, pages 96–103, Barcelona, Spain.

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

P. Pajas and J. Štěpánek. 2008. Recent advances in a feature-rich framework for treebank annotation. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 673–680, Manchester, UK.

M. Palmer, D. Gildea, and P. Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31:71–105.

I. Titov and J. Henderson. 2007. A latent variable model for generative dependency parsing. In *Proceedings of the International Conference on Parsing Technologies (IWPT-07)*, pages 144–155, Prague, Czech Republic.