# Semantic Role Labelling with minimal resources: Experiments with French

**Rasoul Kaljahi**[†‡], **Jennifer Foster**[†], **Johann Roturier**[‡]
[†]NCLT, School of Computing, Dublin City University, Ireland
{rkaljahi, jfoster}@computing.dcu.ie
[‡]Symantec Research Labs, Dublin, Ireland
johann_roturier@symantec.com

## Abstract

This paper describes a series of French semantic role labelling experiments which show that a small set of manually annotated training data is superior to a much larger set containing semantic role labels which have been projected from a source language via word alignment. Using universal part-of-speech tags and dependencies makes little difference over the original fine-grained tagset and dependency scheme. Moreover, there seems to be no improvement gained from projecting semantic roles between direct translations than between indirect translations.

## 1 Introduction

Semantic role labelling (SRL) (Gildea and Jurafsky, 2002) is the task of identifying the predicates in a sentence, their semantic arguments and the roles these arguments take. The last decade has seen considerable attention paid to statistical SRL, thanks to the existence of two major hand-crafted resources for English, namely, FrameNet (Baker et al., 1998) and PropBank (Palmer et al., 2005). Apart from English, only a few languages have SRL resources and these resources tend to be of limited size compared to the English datasets.

French is one of those languages which suffer from a scarcity of hand-crafted SRL resources. The only available gold-standard resource is a small set of 1000 sentences taken from Europarl (Koehn, 2005) and manually annotated with Propbank verb predicates (van der Plas et al., 2010b). This dataset is then used by van der Plas et al. (2011) to evaluate their approach to projecting the SRLs of English sentences to their translations in French. They additionally build a large, "artificial" or automatically labelled dataset of approximately 1M Europarl sentences by projecting the SRLs from English sentences to their French translations and use it for training an SRL system.

We build on the work of van der Plas et al. (2010b) by answering the following questions: 1) *How much artificial data is needed to train an SRL system?* 2) *Is it better to use direct translations than indirect translations*, i.e. is it better to use for projection a source-target pair where the source represents the original sentence and the target represents its direct translation as opposed to a source-target pair where the source and target are both translations of an original sentence in a third language? 3) *Is it better to use coarse-grained syntactic information (in the form of universal part-of-speech tags and universal syntactic dependencies) than to use fine-grained syntactic information?* We find that SRL performance levels off after only 5K training sentences obtained via projection and that direct translations are no more useful than indirect translations. We also find that it makes very little difference to French SRL performance whether we use universal part-of-speech tags and syntactic dependencies or more fine-grained tags and dependencies.

The surprising result that SRL performance levels off after just 5K training sentences leads us to directly compare the small hand-crafted set of 1K sentences to the larger artificial training set. We use 5-fold cross-validation on the small dataset and find that the SRL performance is substantially higher (>10 $F_1$ in identification and classification) when the hand-crafted annotations are used.

## 2 Related Work

There has been relatively few works in French SRL. Lorenzo and Cerisara (2012) propose a clustering approach for verb predicate and argument labelling (but not identification). They choose

This work is licensed under a Creative Commons Attribution 4.0 International Licence. Page numbers and proceedings footer are added by the organisers. Licence details: http://creativecommons.org/licenses/by/4.0/



*Proceedings of the Third Joint Conference on Lexical and Computational Semantics (*SEM 2014)*, pages 87–92,
Dublin, Ireland, August 23-24 2014.

VerbNet style roles (Schuler, 2006) and manually annotate sentences with them for evaluation, achieving an $F_1$ of 78.5.

Gardent and Cerisara (2010) propose a method for semi-automatically annotating the French dependency treebank (Candito et al., 2010) with Propbank core roles (no adjuncts). They first manually augment TreeLex (Kupść and Abeillé, 2008), a syntactic lexicon of French, with semantic roles of syntactic arguments of verbs (i.e. verb subcategorization). They then project this annotation to verb instances in the dependency trees. They evaluate their approach by performing error analysis on a small sample and suggest directions for improvement. The annotation work is however at its preliminary stage and no data is published.

As mentioned earlier, van der Plas et al. (2011) use word alignments to project the SRLs of the English side of EuroParl to its French side resulting in a large artificial dataset. This idea is based on the *Direct Semantic Transfer* hypothesis which assumes that a semantic relationship between two words in a sentence can be transferred to any two words in the translation which are aligned to these source-side words. Evaluation on their 1K manually-annotated dataset shows that a syntactic-semantic dependency parser trained on this artificial data set performs significantly better than directly projecting the labelling from its English side – a promising result because, in a real-world scenario, the English translations of the French data to be annotated do not necessarily exist.

Padó and Lapata (2009) also make use of word alignments to project SRLs from English to German. The word alignments are used to compute the semantic similarity between syntactic constituents. In order to determine the extent of semantic correspondence between English and German, they manually annotate a set of parallel sentences and find that about 72% of the frames and 92% of the argument roles exist in both sides, ignoring their lexical correspondence.

## 3 Datasets, SRL System and Evaluation

We use the two datasets described in (van der Plas et al., 2011) and the delivery report of the *Classic* project (van der Plas et al., 2010a). These are the gold standard set of 1K sentences which was annotated by manually identifying each verb predicate, finding its equivalent English frameset in PropBank and identifying and labelling its arguments based on the description of the frameset (henceforth known as *Classic1K*), and the synthetic dataset consisting of more than 980K sentences (henceforth known as *Classic980K*), which was created by word aligning an English-French parallel corpus (Europarl) using GIZA++ (Och and Ney, 2003) and projecting the French SRLs from the English SRLs via the word alignments. The joint syntactic-semantic parser described in (Titov et al., 2009) was used to produce the English SRLs and the dependency parses of the French side were produced using the `ISBN` parser described in (Titov and Henderson, 2007).

We use LTH (Björkelund et al., 2009), a dependency-based SRL system, in all of our experiments. This system was among the best-performing systems in the CoNLL 2009 shared task (Hajič et al., 2009) and is straightforward to use. It comes with a set of features tuned for each shared task language (English, German, Japanese, Spanish, Catalan, Czech, Chinese). We compared the performance of the English and Spanish feature sets on French and chose the former due to its higher performance (by 1 $F_1$ point).

To evaluate SRL performance, we use the CoNLL 2009 shared task scoring script[1], which assumes a semantic dependency between the argument and predicate and the predicate and a dummy root node and then calculates the precision (`P`), recall (`R`) and $F_1$ of identification of these dependencies and classification (labelling) of them.

## 4 Experiments

### 4.1 Learning Curve

The ultimate goal of SRL projection is to build a training set which partially compensates for the lack of hand-crafted resources. van der Plas et al. (2011) report encouraging results showing that training on their projected data is beneficial over directly obtaining the annotation via projection which is not always possible. Although the quality of such automatically-generated training data may not be comparable to the manual one, the possibility of building much bigger data sets may provide some advantages. Our first experiment investigates the extent to which the size of the synthetic training set can improve performance.

We randomly select 100K sentences from *Classic980K*, shuffle them and split them into 20 sub-

[1] `https://ufal.mff.cuni.cz/conll2009-st/eval09.pl`

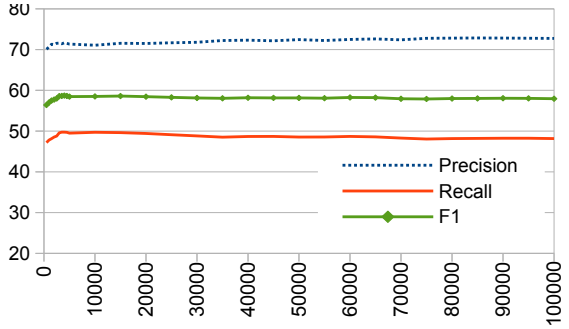


Figure 1: Learning curve with 100K training data of projected annotations

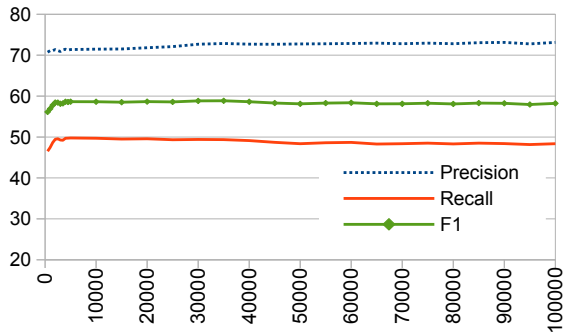


Figure 2: Learning curve with 100K training data of projected annotations on only direct translations

sets of 5K sentences. We then split the first 5K into 10 sets of 500 sentences. We train SRL models on the resulting 29 subsets using LTH. The performance of the models evaluated on *Classic1K* is presented in Fig. 1. Surprisingly, the best $F_1$ (58.7) is achieved by only 4K sentences, and after that the recall (and consequently $F_1$) tends to drop though precision shows a positive trend, suggesting that the additional sentences bring little information. The large gap between precision and recall is also interesting, showing that the projections do not have wide semantic role coverage.[2]

### 4.2 Direct Translations

Each sentence in Europarl was written in one of the official languages of the European Parliament and translated to all of the other languages. Therefore both sides of a parallel sentence pair can be indirect translations of each other. van der Plas et al. (2011) suggest that translation divergence may affect automatic projection of semantic roles. They therefore select for their experiments only those 276K sentences from the 980K which are direct translations between English and French. Motivated by this idea, we replicate the learning curve in Fig. 1 with another set of 100K sentences randomly selected from only the direct translations. The curve is shown in Fig. 2. There is no noticeable difference between this and the graph in Fig. 1, suggesting that the projections obtained via direct translations are not of higher quality.

### 4.3 Impact of Syntactic Annotation

Being a dependency-based semantic role labeller, LTH employs a large set of features based on syntactic dependency structure. This inspires us to compare the impact of different types of syntactic annotations on the performance of this system.

Based on the observations from the previous sections, we choose two different sizes of training sets. The first set contains the first 5K sentences from the original 100K, as we saw that more than this amount tends to diminish performance. The second set contains the first 50K from the original 100K, the purpose of which is to check if changing the parses affects the usefulness of adding more data. We will call these data sets *Classic5K* and *Classic50K* respectively.

Petrov et al. (2012) create a set of 12 universal part-of-speech (POS) tags which should in theory be applicable to any natural language. It is interesting to know whether these POS tags are more useful for SRL than the original set of the 29 more fine-grained POS tags used in French Treebank which we have used so far. To this end, we convert the original POS tags of the data to universal POS tags and retrain and evaluate the SRL models. The results are given in the second row of Table 1 (`OrgDep+UniPOS`). The first row of the table (`Original`) shows the performance using the original annotation. Even though the scores increase in most cases – due mostly to a rise in recall – the changes are small. It is worth noting that identification seems to benefit more from the universal POS tags.

Similar to universal POS tags, McDonald et al. (2013) introduce a set of 40 universal dependency types which generalize over the dependency structure specific to several languages. For French, they provide a new treebank, called `uni-dep-tb`, manually annotating 16,422 sentences from vari-

[2]Note that our results are not directly comparable with (van der Plas et al., 2011) because they split *Classic1K* into development and test sets, while we use the whole set for testing. We do not have access to their split.

| | 5K | | | | | | 50K | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Identification | | | Classification | | | Identification | | | Classification | | |
| | P | R | $F_1$ | P | R | $F_1$ | P | R | $F_1$ | P | R | $F_1$ |
| Original | 85.95 | 59.64 | 70.42 | **71.34** | 49.50 | **58.45** | 86.67 | 58.07 | 69.54 | **72.44** | 48.54 | 58.13 |
| OrgDep+UniPOS | **86.71** | **60.46** | **71.24** | 71.11 | **49.58** | 58.43 | **86.82** | 58.71 | **70.05** | 72.30 | **48.90** | **58.34** |
| StdUniDep+UniPOS | 86.14 | 59.76 | 70.57 | 70.60 | 48.98 | 57.84 | 86.38 | **58.90** | 70.04 | 71.61 | 48.83 | 58.07 |
| CHUniDep+UniPOS | 85.98 | 59.21 | 70.13 | 70.66 | 48.66 | 57.63 | 86.47 | 58.26 | 69.61 | 71.74 | 48.34 | 57.76 |

Table 1: SRL performance using different syntactic parses with Classic 5K and 50K training sets

ous domains. We now explore the utility of this new dependency scheme in SRL.

The French universal dependency treebank comes in two versions, the first using the standard dependency structure based on basic Stanford dependencies (de Marneffe and Manning, 2008) where content words are the heads except in copula and adposition constructions, and the second which treats content words as the heads for all constructions without exemption. We use both schemes in order to verify their effect on SRL.

In order to obtain universal dependencies for our data, we train parsing models with MaltParser (Nivre et al., 2006) using the entire uni-dep-tb.[3] We then parse our data using these MaltParser models. The input POS tags to the parser are the universal POS tags used in OrgDep+UniPOS. We train and evaluate new SRL models on these data. The results are shown in the third and fourth rows of Table 1. StdUniDept+UniPOS is the setting using standard dependencies and CHUDep+UPOS using content-head dependencies.

According to the third and fourth rows in Table 1, content-head dependencies are slightly less useful than standard dependencies. The general effect of universal dependencies can be compared to those of original ones by comparing these results to OrgDep+UniPOS - the use of universal dependencies appears to have only a modest (negative) effect. However, we must be careful of drawing too many conclusions because in addition to the difference in dependency schemes, the training data used to train the parsers as well as the parsers themselves are different.

Overall, we observe that the universal annotations can be reliably used when the fine-grained annotation is not available. This can be especially useful for languages which lack such resources and require techniques such as cross-lingual transfer to replace them.

[3]Based on our preliminary experiments on the parsing performance, we use LIBSVM as learning algorithm, nivreeager as parsing algorithm for the standard dependency models and stackproj for the content-head ones.

| | Identification | | | Classification | | |
|---|---|---|---|---|---|---|
| | P | R | $F_1$ | P | R | $F_1$ |
| 1K | 83.76 | 83.00 | 83.37 | 68.40 | 67.78 | **68.09** |
| 5K | **85.94** | 59.62 | 70.39 | 71.30 | 49.47 | 58.40 |
| 1K+5K | 85.74 | 66.53 | 74.92 | **71.48** | 55.46 | 62.46 |
| SelfT | 83.82 | **83.66** | **83.73** | 67.91 | **67.79** | 67.85 |

Table 2: Average scores of 5-fold cross-validation with Classic 1K (1K), 5K (5K), 1K plus 5K (1K+5K) and self-training with 1K seed and 5K unlabeled data (SelfT)

### 4.4 Quality vs. Quantity

In Section 4.1, we saw that adding more data annotated through projection did not elevate SRL performance. In other words, the same performance was achieved using only a small amount of data. This is contrary to the motivation for creating synthetic training data, especially when the hand-annotated data already exist, albeit in a small size. In this section, we compare the performance of SRL models trained using manually-annotated data with SRL models trained using 5K of artificial or synthetic training data. We use the original syntactic annotations for both datasets.

To this end, we carry out a 5-fold cross-validation on *Classic1K*. We then evaluate the *Classic5K* model, on each of the 5 test sets generated in the cross-validation. The average scores of the two evaluation setups are compared. The results are shown in Table 2.

While the 5K model achieves higher precision, its recall is far lower resulting in dramatically lower $F_1$. This high precision and low recall is due to the low confidence of the model trained on projected data suggesting that a considerable amount of information is not transferred during the projection. This issue can be attributed to the fact that the

Classic projection uses intersection of alignments in the two translation directions, which is the most restrictive setting and leaves many source predicates and arguments unaligned.

We next add the *Classic5K* projected data to the manually annotated training data in each fold of another cross-validation setting and evaluate the resulting models on the same test sets. The results are reported in the third row of the Table 2 (1K+5K). As can be seen, the low quality of the projected data significantly degrades the performance compared to when only manually-annotated data are used for training.

Finally, based on the observation that the quality of labelling using manually annotated data is higher than using the automatically projected data, we replicate 1K+5K with the 5K data labelled using the model trained on the training subset of 1K at each cross-validation fold. In other words, we perform a one-round self-training with this model. The performance of the resulting model evaluated in the same cross-validation setting is given in the last row of Table 2 (SelfT).

As expected, the labelling obtained by models trained on manual annotation are more useful than the projected ones when used for training new models. It is worth noting that, unlike with the 1K+5K setting, the balance between precision and recall follows that of the 1K model. In addition, some of the scores are the highest among all results, although the differences are not significant.

### 4.5 How little is too little?

In the previous section we saw that using a manually annotated dataset with as few as 800 sentences resulted in significantly better SRL performance than using projected annotation with as many as 5K sentences. This unfortunately indicates the need for human labour in creating such resources. It is interesting however to know the lower bound of this requirement. To this end, we reverse our cross-validation setting and train on 200 and test on 800 sentences. We then compare to the 5K models evaluated on the same 800 sentence sets at each fold. The results are presented in Table 3. Even with only 200 manually annotated sentences, the performance is considerably higher than with 5K sentences of projected annotations. However, as one might expect, compared to when 800 sentences are used for training, this small model performs significantly worse.

| | Identification | | | Classification | | |
|---|---|---|---|---|---|---|
| | P | R | $F_1$ | P | R | $F_1$ |
| 1K | 82.34 | **79.61** | **80.95** | 64.14 | **62.01** | **63.06** |
| 5K | **85.95** | 59.64 | 70.42 | **71.34** | 49.50 | 58.45 |

Table 3: Average scores of 5-fold cross-validation with Classic 1K (1K) and 5K (5K) using 200 sentences for training and 800 for testing at each fold

## 5 Conclusion

We have explored the projection-based approach to SRL by carrying out experiments with a large set of French semantic role labels which have been automatically transferred from English. We have found that increasing the number of these artificial projections that are used in training an SRL system does not improve performance as might have been expected when creating such a resource. Instead it is better to train directly on what little gold standard data is available, even if this dataset contains only 200 sentences. We suspect that the disappointing performance of the projected dataset originates in the restrictive way the word alignments have been extracted. Only those alignments that are in the intersection of the English-French and French-English word alignment sets are retained resulting in low SRL recall. Recent preliminary experiments show that less restrictive alignment extraction strategies including extracting the union of the two sets or source-to-target alignments lead to a better recall and consequently $F_1$ both when used for direct projection to the test data or for creating the training data and then applying the resulting model to the test data.

We have compared the use of universal POS tags and dependency labels to the original, more fine-grained sets and shown that there is only a little difference. However, it remains to be seen whether this finding holds for other languages or whether it will still hold for French when SRL performance can be improved. It might also be interesting to explore the combination of universal dependencies with fine-grained POS tags.

## Acknowledgments

This research has been supported by the Irish Research Council Enterprise Partnership Scheme (EPSPG/2011/102) and the computing infrastructure of the CNGL at DCU. We thank Lonneke van der Plas for providing us the Classic data. We also thank the reviewers for their helpful comments.

## References


Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The berkeley framenet project. In *Proceedings of the 36th ACL*, pages 86–90.

Anders Björkelund, Love Hafdell, and Pierre Nugues. 2009. Multilingual semantic role labeling. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*, pages 43–48.

Marie Candito, Benot Crabbé, and Pascal Denis. 2010. Statistical french dependency parsing: treebank conversion and first results. In *Proceedings of LREC'2010*.

Marie-Catherine de Marneffe and Christopher D. Manning. 2008. The stanford typed dependencies representation. In *Proceedings of the COLING Workshop on Cross-Framework and Cross-Domain Parser Evaluation*.

Claire Gardent and Christophe Cerisara. 2010. Semi-Automatic Propbanking for French. In *TLT9 - The Ninth International Workshop on Treebanks and Linguistic Theories*.

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

Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, Pavel Straňák, Mihai Surdeanu, Nianwen Xue, and Yi Zhang. 2009. The conll-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning: Shared Task*, pages 1–18.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86.

Anna Kupść and Anne Abeillé. 2008. Growing treelex. In *Proceedings of the 9th International Conference on Computational Linguistics and Intelligent Text Processing*, CICLing'08, pages 28–39.

Alejandra Lorenzo and Christophe Cerisara. 2012. Unsupervised frame based semantic role induction: application to french and english. In *Proceedings of the ACL 2012 Joint Workshop on Statistical Parsing and Semantic Processing of Morphologically Rich Languages*, pages 30–35.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97.

Joakim Nivre, Johan Hall, and Jens Nilsson. 2006. Maltparser: A data-driven parser-generator for dependency parsing. In *In Proceedings of LREC*.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51.

Sebastian Padó and Mirella Lapata. 2009. Cross-lingual annotation projection of semantic roles. *J. Artif. Int. Res.*, 36(1):307–340.

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

Slav Petrov, Dipanjan Das, and Ryan McDonald. 2012. A universal part-of-speech tagset. In *Proceedings of LREC*, May.

Karin Kipper Schuler. 2006. *VerbNet: A Broad-Coverage, Comprehensive Verb Lexicon*. Ph.D. thesis, University of Pennsylvania.

Ivan Titov and James Henderson. 2007. A latent variable model for generative dependency parsing. In *Proceedings of the 10th International Conference on Parsing Technologies*, pages 144–155.

Ivan Titov, James Henderson, Paola Merlo, and Gabriele Musillo. 2009. Online projectivisation for synchronous parsing of semantic and syntactic dependencies. In *In Proceedings of the Internation Joint Conference on Artificial Intelligence (IJCAI)*, pages 1562–1567.

Lonneke van der Plas, James Henderson, and Paola Merlo. 2010a. D6. 2: Semantic role annotation of a french-english corpus.

Lonneke van der Plas, Tanja Samardžić, and Paola Merlo. 2010b. Cross-lingual validity of propbank in the manual annotation of french. In *Proceedings of the Fourth Linguistic Annotation Workshop*, LAW IV '10, pages 113–117.

Lonneke van der Plas, Paola Merlo, and James Henderson. 2011. Scaling up automatic cross-lingual semantic role annotation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 299–304.