

# X-SRL: A Parallel Cross-Lingual Semantic Role Labeling Dataset

Angel Daza<sup>†</sup> ♣ and Anette Frank ♣

Leibniz ScienceCampus “Empirical Linguistics and Computational Language Modeling”

♣Heidelberg University / <sup>†</sup>Institut für Deutsche Sprache Mannheim

Germany

{daza, frank}@cl.uni-heidelberg.de

## Abstract

Even though SRL is researched for many languages, major improvements have mostly been obtained for English, for which more resources are available. In fact, existing multilingual SRL datasets contain disparate annotation styles or come from different domains, hampering generalization in multilingual learning. In this work we propose a method to automatically construct an SRL corpus that is *parallel in four languages: English, French, German, Spanish*, with unified predicate and role annotations that are fully comparable across languages. We apply high-quality machine translation to the English CoNLL-09 dataset and use multilingual BERT to project its high-quality annotations to the target languages. We include human-validated test sets that we use to measure the projection quality, and show that projection is denser and more precise than a strong baseline. Finally, we train different SOTA models on our novel corpus for mono- and multilingual SRL, showing that the multilingual annotations improve performance especially for the weaker languages.

## 1 Introduction

Semantic Role Labeling (SRL) is the task of extracting semantic predicate-argument structures from sentences. One of the most widely used labeling schemes for this task is based on PropBank (Palmer et al., 2005). It comes in two variants: span-based labeling, where arguments are characterized as word-spans (Carreras and Màrquez, 2005; Pradhan et al., 2012), and head-based labeling, which only labels the syntactic head (Hajič et al., 2009). In this work we focus on head-based labeling, as it is applied in the multilingual CoNLL-09 shared task dataset, comprising 7 languages.

The performance of English SRL has considerably improved in recent years through continuous

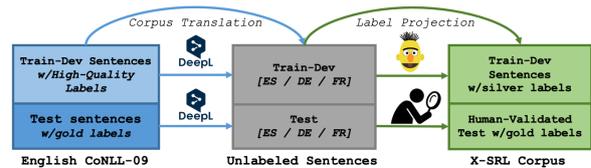


Figure 1: Method to create X-SRL. We automatically translate the English CoNLL-09 corpus, use a fast label projection method for *train-dev* and get human annotators to select the appropriate head words on the target sentences to obtain gold annotations for the *test* sets.

refinements of Deep Neural Network (DNN) models (Zhou and Xu, 2015; He et al., 2017; Marcheggiani et al., 2017; Cai et al., 2018); however, although the CoNLL-09 SRL dataset already covers 7 languages, other languages have not received the same level of attention. This situation may be due to factors such as i) the lack of sufficient training data to successfully apply a language-agnostic DNN model; ii) the fact that creating new SRL datasets is resource-consuming; iii) current label projection methods suffering from low recall; finally, iv) even in cases where annotated resources are available in other languages, often they were automatically converted from independent pre-existing annotation schemes or labeled with automatic methods, resulting in data quality and labeling schema divergences, hampering the effectiveness of unified models that can be applied in multilingual settings.

In this paper we offer a **multilingual parallel SRL corpus – X-SRL** – for English (EN), German (DE), French (FR) and Spanish (ES) that is based on English gold annotations and *shares the same labeling scheme across languages*.<sup>1</sup> Our corpus has two major advantages compared to existing datasets: first, since it is a *parallel corpus*, all sen-

<sup>1</sup>[https://github.com/Heidelberg-NLP/xsrl\\_mbert\\_aligner](https://github.com/Heidelberg-NLP/xsrl_mbert_aligner)

tences are semantically equivalent allowing us to analyze performance at the sentence-level and to better understand the reasons for SRL score divergences across languages<sup>2</sup>; second, we expect that models trained jointly on multiple languages annotated with a homogeneous labeling style will be able to better generalize across languages<sup>3</sup>. Moreover, by minimizing the need of specialized human annotators, our parallel corpus construction method is *lightweight and portable*, since it is built on three main components: i) A high-quality annotated dataset in the source language, in this case the English CoNLL-09 corpus (Hajič et al., 2009), ii) a high-quality SOTA Machine Translation system, we are using DeepL Pro<sup>4</sup>; and iii) multilingual contextual word representations, in this case multilingual BERT (mBERT) (Devlin et al., 2019). The situation for these multilingual resources is improving with each day, and thus our method, in perspective, could be followed for producing training data for more lower-resource languages. Importantly, although we automatically project labels from English to the newly available corpora in the different languages for *train* and *dev* sections, we also provide *test* sets on which humans assess the quality of the automatic translations and select the valid predicates as well as the appropriate target head words for the target sentences. Having a human-validated test set ensures solid benchmarking for SRL systems, and additionally allows us to assess the validity of the proposed automatic projection for the rest of the data.

Our projection method works as follows (see also Figure 1): We obtain automatic translations of the English CoNLL-09 corpus into each of our target languages; then, we automatically label them without applying any language-pair specific label projection model, but use mBERT with additional filters as a means for alignment. We show that by following this approach we obtain a *more densely annotated dataset* compared to an existing SOTA label projection method (Akbik and Vollgraf, 2018). In short, our contributions are:

- The first fully parallel SRL dataset with

<sup>2</sup>E.g., German F1 score on the CoNLL-09 dataset lags 10 points behind English, but with currently available datasets we cannot be sure if this is due to differences in the available training data or because of language-specific characteristics.

<sup>3</sup>It is not straightforward to use the CoNLL-09 data in a multilingual model: for example, annotations for German use a role inventory with roles A0-A9, and a one-to-one mapping to all English labels is not available.

<sup>4</sup><https://www.deepl.com/translator>

dense, homogeneous annotations and human-validated test sets covering four languages: English, French, German and Spanish.

- A simple but effective novel method to project existing SRL annotations from English to lower-resource languages.
- A fast method to create a human-supervised test set that allows us to explore the syntactic and semantic divergences in SRL across languages and to assess performance differences.
- We provide quality measures for our projection based on the human validation process.
- We demonstrate the multilingual generalization capabilities of our corpus by training different SOTA baseline models on our dataset.

## 2 Related Work

Semi-automatic annotation projection has been applied to different SRL frameworks. Pado (2007); Padó and Lapata (2009) proposed a projection method for FrameNet (Baker et al., 1998) semantic roles that searches for the best alignment of source and target constituent trees, and also created a small human-validated test set for benchmarking.

A number of PropBank resources are available for different languages: the benchmark datasets CoNLL-09 (Hajič et al., 2009) and CoNLL-12 (Pradhan et al., 2012) are well-established, however, a direct cross-lingual comparison of SRL performance across the covered languages is not possible. The reason being that the language-specific datasets come from different sources and were not conceived for such a comparison.

On the other hand, van der Plas et al. (2010) attested the validity of English PropBank labels for French and directly applied them on French data. This motivated SRL projection methods such as van der Plas et al. (2011) and Akbik et al. (2015), which aim to generate a common label set across languages. A known issue with this approach is the need for good quality parallel and sentence-level filtered data. For this reason they used existing parallel corpora, Europarl (Koehn, 2005) and UN (Ziemski et al., 2016), automatically labeled the English side with SRL annotations and transferred them to the corresponding translations. The major issue with this is that evaluation against ground-truth and detailed error analysis on the target languages are not possible, since all annotations are

automatic and come from noisy sources. Likewise, the Universal Proposition Bank<sup>5</sup> (Akbik et al., 2015; Akbik and Li, 2016), adds an SRL layer on top of the Universal Dependencies (de Marneffe et al., 2014) corpus, which covers eight different languages. However, i) the original corpora come from independent sources and are not parallel, ii) the source sentences were automatically labeled containing noise even before the alignment step, and iii) the test sets also contain automatic projections without human validation of the labels.

In contrast, we present a corpus that transfers English high-quality labels to the target side, thus projecting the same labeling style to the other languages; more importantly, we conceive of this corpus, from the very beginning, as a parallel resource with translation equivalence with the source and target languages at the sentence-level. In addition, we create a human-validated test set to allow for proper benchmarking in each language.

The use of synthetic data generated by automatic translation has proven to improve performance for MT (Sennrich et al., 2016) and Argumentation Mining (Eger et al., 2018). We similarly create a parallel corpus using automatic translation, however, to our knowledge, we are the first to create a directly comparable multilingual SRL dataset using automatic translation with a manually validated test set, minimizing human labour.

Another attempt to close the gap between languages is by training multilingual SRL systems. He et al. (2019) propose a biaffine scorer with syntax rules to prune of candidates, achieving SOTA independently in all languages from CoNLL-09. Mulcaire et al. (2018) and Daza and Frank (2019) train a single model using input data from different languages and obtain modest improvements, especially for languages where less monolingual training data is available. In this sense, our X-SRL corpus contributes with more compatible training data across languages, and aims to improve the performance of jointly trained multilingual models.

### 3 Building X-SRL

In this section we first explain our method for translating the English CoNLL-09 SRL dataset (§3.1) into three target languages (DE, ES, FR)<sup>6</sup>. In §3.2

<sup>5</sup><https://github.com/System-T/UniversalPropositions>

<sup>6</sup>We chose these languages given the availability of annotators to validate the quality of test set translations. We hope that future work will apply our method to further languages.

we describe how the human-validated labels (only for the test sets) were obtained in an efficient way, and report annotator agreement statistics in §3.3. The details of how we perform (automatic) label projection enhanced with filtering for *train/dev* are given in §3.4. With this we achieve big annotated SRL datasets for three new languages (cf. Table 3).

When building the X-SRL dataset, in line with the current PropBank SRL data available in different languages, we focus on verbal predicates only. Note that the English CoNLL-09 data includes both verbal and nominal predicate annotations, yet this is due to the NomBank project (Meyers et al., 2004) being available for that language. By contrast, the remaining languages with PropBank SRL training data (including the CoNLL-09 non-English data) only provide annotations for verbal predicates. While we could attempt projecting the English nominal predicate annotations and create an X-SRL dataset that includes nominal SRL for all target languages – which would mean a big advance over the current situation – admitting nominal and verbal SRL annotations in a multilingual setting would confront us with many translation shifts. We could try to capture these for the manually curated test set, however we would run a risk of generating noisy or scarce target annotations when projecting them for the *train/dev* sections.<sup>7</sup>

#### 3.1 Dataset Translation

We aim to produce high-quality labeled corpora while reducing as much as possible the amount of time, cost and human intervention needed to fulfill this task. We use Machine Translation to perform dataset translation, obviating the need of human translator services. As previous work (Tiedemann and Agic, 2016; Tyers et al., 2018) has shown, automatic translations are useful as supervision for syntactic dependency labeling tasks since they are quite close to the source languages; likewise, in Argumentation Mining, Eger et al. (2018) achieve comparable results to using human-translated data. One could argue that by automatically translating the English source, we could run into a problem of

<sup>7</sup>The reasons are complex: First, by including nominal SRL, we would be confronted with translation shifts in both directions, e.g. N-to-V or V-to-N translations. For these, we'd have to verify whether they correspond to valid verbalizations or nominalizations on the target side. This would lead to considerable overhead and, most likely, noise in automatic projection. Also, translation shifts often involve light verb constructions, which require special role annotations. These would be difficult to assign in automatic projection. We thus defer the inclusion of nominal SRL to future work.

(1)	a.	People aren't <b>panicking</b> .
	b.	La gente no está <b>entrando en pánico</b> . The people not are entered in panic.
(2)	a.	The account had <b>billed</b> about \$6 million in 1988, according to Leading National Advertisers.
	b.	Das Konto hatte 1988 etwa 6 Millionen Dollar <b>in Rechnung gestellt</b> , so die Leading National Advertisers. The account had 1988 about 6 million dollars in invoice put, so the Leading National Advertisers.
(3)	a.	The economy does, however, <b>depend</b> on the confidence of businesses, consumers and foreign investors.
	b.	Die Wirtschaft <b>hängt</b> jedoch vom Vertrauen von Unternehmen, Verbrauchern und ausländischen Investoren <b>ab</b> . The economy hangs however from-the confidence of businesses, consumers and foreign investors off.
(4)	a.	But while the <b>New York Stock Exchange</b> did n't fall apart Friday as the <b>Dow Jones Industrial Average</b> plunged 190.58 points.
	b.	Mais si la <b>Bourse de New York</b> ne s' est pas effondrée vendredi alors que le <b>Dow Jones Industrial Average</b> a chuté de 190.58 points. But if the Exchange of New York not Refl is not collapsed Friday when that the Dow Jones Industrial Average has fallen by 190.58 points.

Table 1: Examples of translation shifts: (1) predicate nominalization on the target side, (2) and (1) source verb converted to a light verb construction on the target side, (3) a source predicate translates to a verb with separable prefix, and (4) instances of Named Entities being translated or not to the target language.

*translationese*.<sup>8</sup> While it would be interesting to study possible shining-through effects in our automatically translated target texts and any potential impact on SRL performance (e.g. by comparing a natural vs. translated test set), our main concern is to preserve the relevant *predicate-argument structures* in order to give a strong-enough signal to train our SRL systems, and our initial assumption relies on the evidence from the mentioned previous works (confirmed by our results) that obtaining relevant training data is possible with MT generated sentences.

We take as source the set of sentences in the English CoNLL-09 dataset, which are tokenized and annotated for part-of-speech (POS), syntactic dependencies, predicate senses and semantic roles. We use DeepL to obtain translations of each sentence into the three target languages. For all target sentences we use spaCy<sup>9</sup> to tokenize, assign POS tags and syntactic dependency annotations. This gives us a 4-way parallel corpus with syntactic information on both sides.

### 3.2 Test Set Annotation

**Annotation Setup.** To confirm the good quality of the translations delivered by DeepL, we hired 12 annotators with a background in translation studies and experience in  $EN \rightarrow T$  translation (we hired 4 annotators for each language pair) to *rate and validate the automatic translations of the test set*<sup>10</sup>

<sup>8</sup>Translationese occurs when – in an attempt to reproduce the meaning of a text in a foreign language – the resulting translation is grammatically correct but carries over language-specific constructs from the source language to the target

<sup>9</sup><https://github.com/explosion/spaCy>

<sup>10</sup>Note that validating a translation that already exists is considerably faster than generating translations from scratch,

by following a guideline that explains the quality validation and the annotation processes<sup>11</sup>. First, we ask them to rate the translations on a scale from 1-5 (worst to best). On the basis of the obtained ratings, we apply a filter and keep only the sentences with *quality rating 3, 4, or 5*. Only on this subset of sentences we require them to do three more tasks: i) we show them the *labeled verbal predicates*<sup>12</sup> in the English sentence and ask them to mark on the target side the words that express the same meaning, ii) we show them a list of *key arguments* (which correspond to the labeled syntactic heads in the English sentence) and likewise, ask them to mark on the target side the expression that best matches each key argument's meaning (marking several words is allowed), and finally iii) we ask them to *fix minor translation mistakes* in order to better reflect the source meaning. Importantly, we ask annotators to flag as **special cases** any one-to-many mappings, and for predicates, any mapping that aligns a source verb to a non-verbal predicate in the target language. We also give the option to map source heads or predicate words to *NONE* when no relevant corresponding expression in the translated sentence can be found.

**Annotation Agreement.** To approximate the inter-annotator agreement, we gave the first 100 sentences to all annotators of each language pair and compute Krippendorff's alpha<sup>13</sup> on this sub-set of sentences. We obtain  $\alpha_{pred_{DE}}=0.75$ ,  $\alpha_{pred_{ES}}=0.73$ ,  $\alpha_{pred_{FR}}=0.78$  for *predicate* and

therefore annotation time and budget dropped significantly.

<sup>11</sup>See Supplement A for the annotation guideline.

<sup>12</sup>We ignore all source nominal predicates.

<sup>13</sup>We use the NLTK implementation with binary distance to compute the agreement of labels.

$\alpha_{role_{DE}}=0.79$ ,  $\alpha_{role_{ES}}=0.70$ ,  $\alpha_{role_{FR}}=0.79$  for *role labels*. This shows that the *fast* annotation method can be trusted.

**Linguistic Validation.** We run a second annotation round where two annotators with linguistic background re-validate the instances that were flagged as **special cases** by translators during the first round (more concretely, the possible *translation shifts*). Specifically, annotators in this phase decide, for each special case, if the annotated label should be deleted or corrected. The cases could fall into one or more of the following categories<sup>14</sup> (see Table 1 for some examples):

- **Nominalizations:** A verbal expression (predicate) in English is translated to a nominal expression in the Target (see Table 1, examples (1, 2)). Since we restrict our dataset to verbal predicates (see fn. 7) we discourage the annotation of nominal predicates even when they preserve the original sense.
- **Light Verb Constructions:** This is a special case of nominalization on the target side, where a noun that corresponds to a verb in the source language is an argument of a so-called 'light' verb with bleached, often aspectual, meaning. In example (2), the verb *billed* is translated to *in Rechnung gestellt* (literally: 'in invoice put'). According to Bonial et al. (2015), the nominal argument of a light verb needs a special role annotation.<sup>15</sup> Since there is no easy automatic method to figure out the target senses, we leave these cases for future work and do not annotate them here.
- **Separable Verb Prefixes:** In German, specific verbs must split off their prefix in certain constructions, even though this prefix crucially contributes to their meaning. In example (3), the German verb is *abhängen* which means *to depend*, while the verb *hängen* means *to hang*. Since the labeling scheme that we are using only allows us to tag one word as the head, annotators were instructed to pick the truncated stem of the verb, given that the particle is a syntactic dependent of it.

<sup>14</sup>This validation was performed independently, according to the annotators' language expertise. However, the annotators discussed general policies and jointly resolved difficult cases.

<sup>15</sup>The noun projects its predicate-specific role set and in addition includes the governing verb with a role ARGM-LVB.

QUALITY (Q)	EN	DE	ES	FR
5	2,399	718	1,758	1,358
4	0	902	407	463
3	0	593	181	274
2	0	164	46	184
1	0	22	15	119
# Sentences Q >2	2,399	2,213	2,346	2,095
# Kept Predicates Q >2	5,217	4,086	4,376	3,770
# Kept Arguments Q >2	14,156	11,050	10,529	9,854

Table 2: EN shows the original numbers for the English CoNLL-09 corpus. The other three languages show the quality distribution and *predicate* and *role* annotations kept after applying the quality and linguistic filters.

- **Multiword Expressions (MWEs):** A single source word is translated to several target words that constitute a single unit of meaning. The translators were allowed to mark more than one target word if the source word meaning could be mapped to a MWE. For these cases, if they did not fall in any of the previous three categories, and since they were manually aligned for being equivalent in meaning, we transfer the source label to the syntactic head of the marked MWE.
- **Named Entities:** are treated as special cases of MWEs. Some NEs, but not all, are (correctly) translated to the target language, which can result in a change of the argument's head. We see both cases in example (4). When NEs are translated to the target language, we need to select the appropriate head: *Exchange* is the head of the NE in English but *Bourse* should be the head in French. We re-locate the label to the NE's syntactic head on the target side.

The linguistic analysis highlights the importance of providing a human-validated test set – as opposed to relying on automatic projection. While the English labels are considered to be gold standard, their transfer to any target language is not straightforward and must be controlled for the mentioned cases to be considered gold standard on the target side. Accordingly, we also consider filters or refinements for the automatic projection and finally, on the basis of our validated test set, we can evaluate how accurate our automatic projection is.

### 3.3 Test Statistics

Table 2 shows the statistics for the final quality distribution for each of the target language datasets according to the translators' ratings. The final test

sets are composed by all sentences with quality level higher than 2. We observe that after applying this filter, the three languages have roughly similar amounts of good quality sentences (between 87% and 97%) as well as similar density of annotations for both predicate and argument labels. The number of sentences that are completely 4-way parallel is 1,714 (71.45% of the original EN corpus). This confirms the intuition that DeepL generates translations that are faithful to the sources. The number of *special cases* analyzed in the second validation step were 294 (DE), 332 (ES) and 1300 (FR), of which 105, 122 and 173, respectively, were considered to be translation shifts and thus were not considered further.

### 3.4 Automatic Projection

The next step is to find an efficient method to automatically transfer the labels in the *train/dev* portions of the data to the target languages without losing too many gold labels. Contrary to the test set, we cannot perform human validation due to the size of the data; here we are mostly interested in getting automatically *good enough* labels to train models. Usually, label projection methods (Pado, 2007; Padó and Lapata, 2009; van der Plas et al., 2011; Akbik et al., 2015; Aminian et al., 2019) rely on the intersection of *source-to-target* and *target-to-source* word alignments to transfer the labels in the least noisy manner, and this way prefer to have higher precision at the expense of lower recall. Instead, we take a novel approach and rely on the shared space of mBERT embeddings (Devlin et al., 2019). Specifically, we compute pair-wise cosine similarity between source and target tokens and emulate word-alignments according to this measure<sup>16</sup>. We show that using mBERT instead of typical word alignments dramatically improves the recall of the projected annotations, and enhanced with filters, it also achieves high enough precision, resulting in a more densely labeled target side and therefore better quality training data is expected. Additionally, previous works show that BERT contextualized representations are useful for monolingual Word Sense Disambiguation (WSD) tasks (Loureiro and Jorge, 2019; Huang et al., 2019) which lets us assume that we can rely on mBERT to find good word-level alignments across languages.

<sup>16</sup>This is similar to what is done as a first step in BERTScore (Zhang et al., 2020) towards computing a metric for (semantic) sentence similarity, but here we use the token-wise similarity as a guide for cross-lingual word alignments.

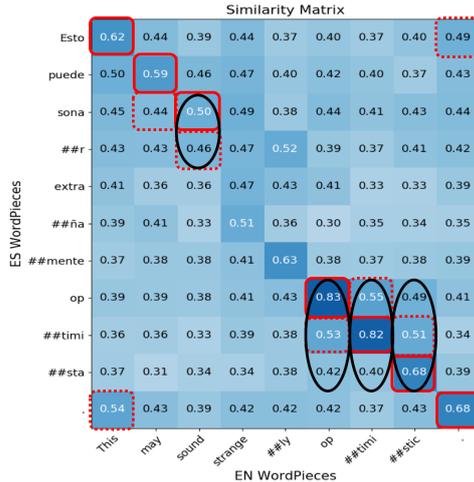


Figure 2: We compute a pair-wise cosine similarity matrix to simulate word alignments. For each column, we look only at source word-pieces with an associated label and keep the top-k ( $k=2$ ) most similar target-side word piece candidates (red squares). The black circles show the aligned full-word.

**BERT Cosine Similarity.** We start with our word tokenized parallel source  $\mathbf{S} = (w_{s_0}, \dots, w_{s_n})$  and target  $\mathbf{T} = (w_{t_0}, \dots, w_{t_m})$  sentences. Then, we use the mBERT tokenizer to obtain word-pieces and their corresponding vectors  $\mathbf{S}' = (v_{s_0}, \dots, v_{s_p})$  and  $\mathbf{T}' = (v_{t_0}, \dots, v_{t_q})$  respectively, where we have  $p$  source word-pieces and  $q$  target word-pieces. We compute the pairwise word-piece cosine similarity between  $\mathbf{S}'$  and  $\mathbf{T}'$ . The cosine similarity between a source word-piece vector and a target word-piece vector is  $\frac{v_{s_i}^T v_{t_j}}{\|v_{s_i}\| \|v_{t_j}\|}$ <sup>17</sup>. The result is a similarity matrix  $\mathbf{SM}$  with  $p$  (columns) and  $q$  (rows) word-pieces (see Figure 2). In addition, we keep a mapping  $\mathbf{S}' \rightarrow \mathbf{S}$  and  $\mathbf{T}' \rightarrow \mathbf{T}$  from each of the word-piece vectors to their original respective word tokens to recover the full-word alignments when needed.

**Word Alignments.** For each column in  $\mathbf{SM}$ , we choose the  $k$  most similar pairs  $(v_s, v_t)$ <sup>18</sup>. This is analogous to a  $\mathcal{A}_{S' \rightarrow T'}$  alignment<sup>19</sup>. The alignment is done from full-word  $w_s$  to full-word  $w_t$ , meaning that for each  $v_s$ , instead of adding a  $v_s \rightarrow v_t$  alignment, we retrieve the full-word  $w_s$  to which  $v_s$  belongs and the  $w_t$  to which  $v_t$  belongs and add a  $w_s \rightarrow w_t$  alignment to the list of candidates for  $w_s$ . At this step, we still permit one-to-many mappings,

<sup>17</sup>We use the implementation of Zhang et al. (2020).

<sup>18</sup> $k$  is a hyperparameter which we chose by hand. The best results were obtained with  $k=2$ .

<sup>19</sup>Conversely, we can simulate a  $\mathcal{A}_{T' \rightarrow S'}$  alignment by defining a similar process for each row in the matrix.

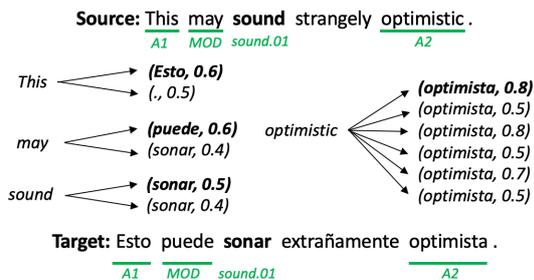


Figure 3: We map word pieces to full-words and apply filters to obtain final source-to-target word alignments.

which means that a  $w_s$  can be associated to more than one  $w_t$  candidates. We retain a dictionary  $D = \{w_s : [(w_{t_1}, sim_{t_1}) \dots (w_{t_x}, sim_{t_x})] | w_s \in S\}$  with their associated similarity scores to keep track of the candidates. See Figure 3 for an example.

**Alignment Modes.** When projecting annotations to the translated training sections, we are confronted with the same *special cases* that we identified in the test set. In the absence of human validation, we have to define filters to eliminate noisy alignments. By only keeping the intersection of alignments  $\mathcal{A}_{S \rightarrow T} \cap \mathcal{A}_{T \rightarrow S}$ , we can get rid of a considerable amount of noisy alignments, however this comes at the cost of a very low recall and a sparsely labeled dataset. Since, we are using an accurate word-similarity measure instead of (noisier) word alignments, we can encourage higher recall by considering all  $\mathcal{A}_{S \rightarrow T}$  alignments and include additional filters to get rid of noisy labels and thus preserve high precision. In (§4.1) we describe in detail the experiments that support this assumption.

**Filtered Projection.** <sup>20</sup> First of all, we eliminate a considerable amount of potential noise by only looking at the  $w_s$ 's that hold a predicate or argument label, while ignoring the rest. Next, for each labeled source predicate, we retrieve from  $D$  the list of target candidates and keep only those that bear a verbal POS tag. If the list contains more than one target candidate we keep the one with the highest score, and if the list is empty we do not project the predicate, as it will most likely instantiate a translation shift or nominalization. Light verbs should be automatically filtered with this method, since the alignment links a verb to a noun and is therefore dropped. For the case of arguments, we also retrieve the candidates from  $D$ . In the ideal case, all candidates belong to the same  $w_t$  and we project the label to that word. Otherwise, we take

<sup>20</sup>[https://github.com/Heidelberg-NLP/xsrl\\_mbert\\_aligner](https://github.com/Heidelberg-NLP/xsrl_mbert_aligner)

the  $w_t$  with more *votes*, i.e. the  $w_t$  that was added most often to the list of candidates. In case of a tie, we turn to the similarity score and transfer the argument label to the  $w_t$  with the highest similarity<sup>21</sup>.

## 4 Experiments and Evaluations

### 4.1 Label Projection

**Intrinsic Evaluation.** Since our test sets are human-validated, we can use them to measure the quality of the label projection methods we have at hand. First, we test the effectiveness of our full method (mBERT+Filters) by comparing it to vanilla cosine similarity (mBERT only) as a projection tool. We apply each method to the test sentences and evaluate the automatically assigned labels against the gold labels provided by annotators. We also show the performance differences when keeping all source to target alignments (S2T) vs. using the intersection of alignments (INTER) when projecting both predicates and arguments. In Table 4 the four combinations can be observed with their specific trade-offs. When using only mBERT with S2T alignments we have high recall but a very mediocre precision; when using INTER alignments we see big gains in precision at the expense of lower recall, as expected. On the other hand, mBERT+Filters obtains consistently better F1, with INTER showing similar behavior to what we observe with the vanilla method, yet with much better precision; however, using full S2T alignments *with filters* gives us the best trade-off: we still achieve around 90% precision and much better recall compared to INTER. This confirms that using S2T alignments (established using mBERT-based cosine similarity) combined with our filters are the best option for projecting labels.

**Extrinsic Evaluation.** Having settled our best method, we compare it with an SRL label projection software: ZAP (Akbik and Vollgraf, 2018)<sup>22</sup>, which also works with the three target languages studied in this paper. ZAP is a pipeline model that takes as input parallel (S, T) sentences, uses source syntactic and semantic parsers to obtain the annotations, and through a trained heuristic word alignment module that uses pre-computed word

<sup>21</sup>Score aggregation would be a straightforward way of computing similarities. However, Zhang et al. (2020) mention that while cosine similarity is good to rank semantic similarity, the computed magnitude is not necessarily proportional, therefore it is not a strict metric. For this reason, we only rely on scores as a decision factor in case of ties.

<sup>22</sup>[www.github.com/zalandoresearch/zap](http://www.github.com/zalandoresearch/zap)

X-SRL	EN			DE			ES			FR		
	Sents	Preds	Args									
Train	39,279	92,908	238,887	39,279	60,861	134,714	39,279	68,844	154,536	39,279	67,878	154,279
Dev	1,334	3,321	8,407	1,334	2,152	4,584	1,334	2,400	5,281	1,334	2,408	5,388
Test	2,399	5,217	14,156	2,213	4,086	11,050	2,346	4,376	10,529	2,095	3,770	9,854

Table 3: Overall statistics for X-SRL.

Method	Lang	INTER			S2T		
		P	R	F1	P	R	F1
mBERT Only	EN-DE	86.6	49.6	63.0	69.0	76.1	72.4
	EN-ES	83.8	68.2	75.2	70.0	84.8	76.7
	EN-FR	82.7	61.8	70.7	67.7	79.5	73.1
mBERT+Filters	EN-DE	96.1	51.8	67.4	92.5	65.8	<b>76.9</b>
	EN-ES	94.0	68.8	79.4	91.9	80.7	<b>85.9</b>
	EN-FR	91.7	63.7	75.2	88.9	74.8	<b>81.2</b>

Table 4: Examining different projection methods on our *human-validated test set*: a) vanilla mBERT cosim (mBERT-Only) vs. adding filters (mBERT+Filters); b) INTER using intersective alignments vs. S2T using full source-to-target alignments. Using S2T alignments and applying filters yield highest F1 alignment score.

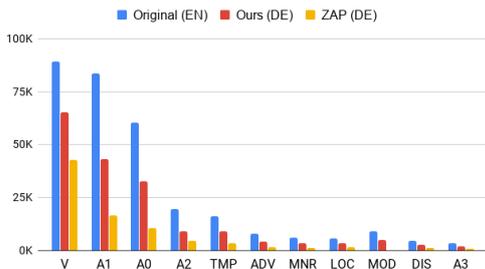


Figure 4: Ten most frequent labels obtained with two label projection methods: OURS vs. ZAP - on the German train set, compared to English source annotations.

translation probabilities, it transfers the labels only when it considers the alignments to be valid, preferring to have fewer, but higher-quality annotations on the target side.

To compare our method to this baseline, we measure the density of the labels on the target training sets after applying both methods to project the labels<sup>23</sup>. Figure 4 shows the case of EN projected to DE where our method consistently recovers more labels from the source, resulting in a more densely annotated training set with comparable label distribution to the EN source. This trend is similar for Spanish and French (overall coverage relative to EN is: DE: 58.9%, ES: 67.3%, FR: 66.9%).

To investigate more deeply why ZAP performs so poorly compared to our method, we use the test

<sup>23</sup>We consider the gold source labels for both methods, thus comparing only their projection performance

	ZAP						OURS					
	PREDICATE			ARGUMENT			PREDICATE			ARGUMENT		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
EN-DE	68.9	15.9	25.9	72.7	15.6	25.7	95.7	76.2	<b>84.9</b>	91.3	61.6	<b>73.6</b>
EN-ES	78.9	34.7	48.2	68.7	30.5	42.2	98.0	89.3	<b>93.4</b>	89.0	76.4	<b>82.2</b>
EN-FR	66.2	21.1	32.0	66.5	24.4	35.7	97.3	85.4	<b>91.0</b>	88.9	69.8	<b>78.2</b>

Table 5: We compare our best projection method with ZAP, a SOTA system for SRL label projection on our test sets. The recall of ZAP is extremely low, damaging their overall scores. In contrast, our method is very good at projecting verbal predicates and arguments.

sets to measure performance. We first evaluate the capacity to transfer source predicates to the target side. Table 5 clearly shows that ZAP fails to transfer many predicates, perhaps because it has unreliable (or no) word-alignment probabilities for infrequent predicates and it is not fine-tuned for this domain (it was trained on Europarl). As a result, also the argument scores are very low, since for each predicate it misses, the system cannot recover any arguments. This highlights the main advantages of our method: by relying on a big multilingual language model i) we obtain high-quality word alignments featuring high precision *and* recall, and ii) we do not need to re-train for other language pairs nor different domains.

## 4.2 Training SRL Systems on X-SRL

At this point we have attested the quality of the automatic method for creating the training sets. Now, as an extrinsic evaluation, we will measure how well can different models learn from our data. To train the models we follow (Zhou and Xu, 2015; He et al., 2017) in the sense that we feed the predicate in training and inference, and we process each sentence as many times as it has predicates, labeling one predicate-argument structure at a time.

**mBERT fine-tuning.** In all settings, we fine-tune mBERT<sup>24</sup>. We use batch size of 16, learning rate of  $5e^{-5}$  and optimize using Adam with weight decay (Loshchilov and Hutter, 2019) and linear schedule with warmup. We train for 5 epochs on our data and pick the epoch that performs best on

<sup>24</sup>We use BertForTokenClassification from <https://huggingface.co/transformers/>

MODEL	EN		DE		ES		FR	
	ZAP	OURS	ZAP	OURS	ZAP	OURS	ZAP	OURS
mBERT EN-tuned	91.0	91.0	69.5	69.5	75.1	75.1	71.9	71.9
mBERT Mono (finetune)	91.0	91.0	58.6	76.1	64.5	80.5	59.5	77.4
mBERT Multi (finetune)	92.4	<b>92.9</b>	63.7	<b>77.0</b>	67.4	<b>81.1</b>	64.1	<b>78.3</b>

Table 6: F1 Score with Fine-tuning mBERT on our training data, created using ZAP vs. OUR projection method and evaluated on our test sets. We compare zero-shot (EN-tuned), mono- and multilingual settings.

MODEL	EN	DE	ES	FR
(Daza and Frank, 2019) Mono	90.9	67.6	56.2	58.1
(Daza and Frank, 2019) Multi	87.6	72.5	77.1	75.2
(Cai et al., 2018) Mono	91.4	76.5	<b>82.6</b>	80.3
(He et al., 2019) Mono	<b>92.4</b>	75.8	82.3	79.3
(He et al., 2019) Multi	92.1	<b>77.3</b>	82.5	<b>80.4</b>

Table 7: F1 Score when training existing SRL models with our data and evaluating on our test. We compare monolingual (Mono) vs using all data available (Multi).

*dev.* Concretely, we explore three settings: The obvious baseline is i) to use only the available English high-quality labels for fine-tuning mBERT and apply zero-shot inference on the other three languages (we call this *EN-tuned*). The other two settings are ii) to fine-tune each language independently with its respective training set (*Mono*) and iii) using all the available data from the four languages to train a single model (*Multi*). Table 6 shows that, as expected, for the *EN-tuned* baseline, English reaches an F1 score of 91, and the other three languages can make good use of mBERT’s knowledge in the zero-shot setting, reaching scores around 70. We also see that our training sets are more complete, obtaining, across the board, higher F1 scores than the training sets projected using ZAP. We observe that training on monolingual data results in improvements for all languages, and finally, the best setting is to use all data at once, improving the already robust mBERT results, and reaching scores of 77, 92, 81 and 78 for DE, EN, ES, FR respectively, about 8 points higher than the zero-shot baseline in the case of German.

**SOTA Models.** Next, we choose three SRL systems that show SOTA results on CoNLL-09 and train them using our data instead. Note that our results are not comparable since our train and test sets are completely different for ES and DE; also the EN results are not comparable since we only label verbal predicates; finally, FR is not present in CoNLL-09. Table 7 summarizes the results. The model of Daza and Frank (2019) is an Encoder-Decoder model that was designed for multilingual

SRL. It performs poorly when trained on monolingual data but improves significantly when trained with more data (multilingual setting). The model of Cai et al. (2018) adapts the biaffine attention scorer of Dozat and Manning (2017) to the SRL task; we note that this model is not designed for handling multilingual data, therefore we only show the monolingual results, which still achieve the best score (82.6) for ES on our test data. Finally, He et al. (2019) generalizes and enhances the biaffine attention scorer with language-specific rules that prune arguments to achieve SOTA on all languages in CoNLL-09. When training this model using our data it achieves the highest scores for EN in the *Mono* setting and for DE and FR when trained with multilingual data. In sum, using our new corpus to train multilingual SRL systems, with SOTA models and finetuning mBERT, we find evidence that the models can use the multilingual annotations for improved performance, especially for the weaker languages.

## 5 Conclusions

In this paper, we present the first fully parallel SRL dataset with homogeneous annotations for four different languages. We included human-validated test sets where we address the linguistic difficulties that emerge when transferring labels across languages – despite transferring gold labels from the source. We introduce and evaluate a novel effective and portable automatic method to transfer SRL labels that relies on the robustness of Machine Translation and multilingual BERT and therefore could be straightforwardly applied to produce SRL data in other languages. Finally, we included an extrinsic evaluation where we train SRL models using our data and obtain consistent results that showcase the generalization capacities emerging from our new 4-way multilingual dataset. Future work should address the application of our method to more and typologically more divergent languages.

## Acknowledgements

We thank our reviewers for their insightful comments. This research was funded by the Leibniz ScienceCampus Empirical Linguistics and Computational Language Modeling, supported by Leibniz Association (grant no. SAS2015-IDS-LWC) and by the Ministry of Science, Research, and Art of Baden-Württemberg. We thank NVIDIA Corporation for a donation of GPUs used in this research.

## References

- Alan Akbik, L.b Chiticariu, M.b Danilevsky, Y.b Li, S.b Vaithyanathan, and H.b Zhu. 2015. [Generating high quality proposition banks for multilingual semantic role labeling](#). *ACL-IJCNLP 2015*, 1:397–418.
- Alan Akbik and Yunyao Li. 2016. [POLYGLOT: multilingual semantic role labeling with unified labels](#). In *Proceedings of ACL-2016 System Demonstrations, Berlin, Germany, August 7-12, 2016*, pages 1–6. Association for Computational Linguistics.
- Alan Akbik and Roland Vollgraf. 2018. [ZAP: An open-source multilingual annotation projection framework](#). In *Proceedings of the 11th Language Resources and Evaluation Conference*, Miyazaki, Japan. European Language Resource Association.
- Maryam Aminian, Mohammad Sadegh Rasooli, and Mona T. Diab. 2019. [Cross-lingual transfer of semantic roles: From raw text to semantic roles](#). In *Proceedings of the 13th International Conference on Computational Semantics, IWCS 2019, Long Papers, Gothenburg, Sweden, May 23-27 May, 2019*, pages 200–210. Association for Computational Linguistics.
- Collin F. Baker, Charles J. Fillmore, John B. Lowe, Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet Project. In *Proceedings of the 17th international conference on Computational linguistics*, volume 1, page 86, Morristown, NJ, USA. Association for Computational Linguistics.
- Claire Bonial, Jena D. Hwang, Julia Bonn, Kathryn Conger, Olga Babko-Malaya, and Martha Palmer. 2015. [:english PropBank annotation guidelines](#). In *Center for Computational Language and Education Research Institute of Cognitive Science University of Colorado at Boulder*.
- Jiaxun Cai, Shexia He, Zuchao Li, and Hai Zhao. 2018. [A full end-to-end semantic role labeler, syntactic-agnostic over syntactic-aware?](#) In *COLING*, pages 2753–2765. Association for Computational Linguistics.
- Xavier Carreras and Lluís Màrquez. 2005. [Introduction to the conll-2005 shared task: Semantic role labeling](#). In *Proceedings of the Ninth Conference on Computational Natural Language Learning, CONLL '05*, pages 152–164, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Angel Daza and Anette Frank. 2019. [Translate and label! an encoder-decoder approach for cross-lingual semantic role labeling](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 603–615, Hong Kong, China. Association for Computational Linguistics.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of deep bidirectional transformers for language understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- Timothy Dozat and Christopher D. Manning. 2017. [Deep biaffine attention for neural dependency parsing](#). In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.
- Steffen Eger, Johannes Daxenberger, Christian Stab, and Iryna Gurevych. 2018. [Cross-lingual argumentation mining: Machine translation \(and a bit of projection\) is all you need!](#) In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 831–844, Santa Fe, New Mexico, USA. Association for Computational Linguistics.
- 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, CoNLL '09*, pages 1–18, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer. 2017. [Deep semantic role labeling: What works and what’s next](#). In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.
- Shexia He, Zuchao Li, and Hai Zhao. 2019. [Syntax-aware multilingual semantic role labeling](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5350–5359, Hong Kong, China. Association for Computational Linguistics.
- Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2019. [GlossBERT: BERT for word sense disambiguation with gloss knowledge](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3509–3514, Hong Kong, China. Association for Computational Linguistics.
- Philipp Koehn. 2005. [Europarl: A Parallel Corpus for Statistical Machine Translation](#).

- Ilya Loshchilov and Frank Hutter. 2019. [Decoupled weight decay regularization](#). In *International Conference on Learning Representations*.
- Daniel Loureiro and Alípio Jorge. 2019. [Language modelling makes sense: Propagating representations through WordNet for full-coverage word sense disambiguation](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5682–5691, Florence, Italy. Association for Computational Linguistics.
- Diego Marcheggiani, Anton Frolov, and Ivan Titov. 2017. [A simple and accurate syntax-agnostic neural model for dependency-based semantic role labeling](#). In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 411–420, Vancouver, Canada. Association for Computational Linguistics.
- Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning. 2014. [Universal Stanford dependencies: A cross-linguistic typology](#). In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 4585–4592, Reykjavik, Iceland. European Language Resources Association (ELRA).
- Adam Meyers, Ruth Reeves, Catherine Macleod, Rachel Szekely, Veronika Zielinska, Brian Young, and Ralph Grishman. 2004. [The NomBank project: An interim report](#). In *Proceedings of the Workshop Frontiers in Corpus Annotation at HLT-NAACL 2004*, pages 24–31, Boston, Massachusetts, USA. Association for Computational Linguistics.
- Phoebe Mulcaire, Swabha Swayamdipta, and Noah A. Smith. 2018. [Polyglot semantic role labeling](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 667–672, Melbourne, Australia. Association for Computational Linguistics.
- Sebastian Pado. 2007. *Cross-lingual Annotation Projection Models for Semantic Role Labeling*. Ph.D. thesis.
- Sebastian Padó and Mirella Lapata. 2009. [Cross-lingual annotation projection of semantic roles](#). *J. Artif. Int. Res.*, 36(1):307–340.
- Martha Palmer, Paul Kingsbury, and Daniel Gildea. 2005. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31.
- Lonneke van der Plas, Paola Merlo, and James Henderson. 2011. Scaling up automatic cross-lingual semantic role annotation. In *ACL (Short Papers)*, pages 299–304. The Association for Computer Linguistics.
- Lonneke van der Plas, Tanja Samardzic, and Paola Merlo. 2010. Cross-lingual validity of propbank in the manual annotation of french. In *Linguistic Annotation Workshop*, pages 113–117. Association for Computational Linguistics.
- Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. [Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes](#). In *Joint Conference on EMNLP and CoNLL - Shared Task*, CoNLL '12, pages 1–40, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. [Improving neural machine translation models with monolingual data](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.
- Jörg Tiedemann and Zeljko Agic. 2016. [Synthetic treebanking for cross-lingual dependency parsing](#). *J. Artif. Intell. Res.*, 55:209–248.
- Francis Tyers, Mariya Sheyanova, Aleksandra Martynova, Pavel Stepachev, and Konstantin Vinogorodskiy. 2018. [Multi-source synthetic treebank creation for improved cross-lingual dependency parsing](#). In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 144–150, Brussels, Belgium. Association for Computational Linguistics.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [BERTScore: Evaluating Text Generation with BERT](#). In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.
- Jie Zhou and Wei Xu. 2015. End-to-end learning of semantic role labeling using recurrent neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1127–1137, Beijing, China. Association for Computational Linguistics.
- Michał Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The united nations parallel corpus v1.0. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France. European Language Resources Association (ELRA).