Improving Multilingual Frame Identification by Estimating Frame Transferability

Jennifer Sikos, *University of Stuttgart*
Michael Roth, *University of Stuttgart*
Sebastian Padó, *University of Stuttgart*

Abstract

A recent research direction in computational linguistics involves efforts to make the field, which used to focus primarily on English, more multilingual and inclusive. However, resource creation often remains a bottleneck for many languages, in particular at the semantic level.

In this article, we consider the case of frame semantic annotation. We investigate how to perform frame selection for annotation in a *target language* by taking advantage of existing annotations in different, *supplementary languages*, with the goal of reducing the required annotation effort in the target language. We measure success by training and testing frame identification models for the target language. We base our selection methods on measuring *frame transferability* in the supplementary language, where we estimate which frames will transfer poorly, and therefore should receive more annotation, in the target language.

We apply our approach on English, German, and French – three languages which have annotations that are similar in size as well as frames with overlapping lexicographic definitions. We find that transferability is indeed a useful indicator and supports a setup where a limited amount of target language data is sufficient to train frame identification systems.

1 Introduction

Semantic frames are structured representations of everyday scenarios or scenes that can be evoked by several predicates (Fillmore, 1982); for example, predicates such as *beat, trounce, demolish,* or *prevail* all evoke a frame about a victor winning over a competitor (BEAT_OPPONENT). Linguistically, frames are scenes that might be realized in different ways. Because of this, semantic frames can be used to account for paraphrase relations among sentences that refer to a shared scenario (*He prevailed over the reigning champ* ≈ *He beat the reigning champ*, Ellsworth and Janin (2007)) or to draw inferences (Ben Aharon et al. (2010)). The Berkeley FrameNet resource for English (Fillmore and Baker, 2001) provides a dictionary of frames where the main components of a frame, including its predicates and semantic roles, are defined. Along with its dictionary, FrameNet provides annotations of frames in text which demonstrate how the frame is used in language.

Frame semantics is also an appealing framework for cross-lingual research, as many frames are thought to be applicable across languages (Boas, 2005). This premise has fueled linguistic research into the applicability of frame semantics to other languages, which have been as varied as German, Spanish, Latvian, Chinese, and Japanese (Gilardi and Baker, 2018). Unfortunately, a recurring bottleneck in these efforts is the need to create frame-semantic annotation. Experiences from existing FrameNet projects show that the timeline for the development of such resources is most likely on the order of years rather than months. This is particularly true for applications of frame semantics in NLP, which involve training *frame-semantic parsers* (e.g., Das et al., 2014, Roth and Lapata, 2015) which require substantial amounts of annotation for each frame.

In this article, we focus on a subproblem of frame-semantic parsing, namely models of frame identification. Frame identification is a disambiguation task where each occurrence of a predicate in context has to be assigned its correct frame given several possible frame candidates. For example, the predicate *cover* can refer to a physical covering (FILLING: *The lid covers the pot*) or to the topic of a communication act (TOPIC: *The textbook covered modality in detail*). The goal of a frame identification system is to take a new instance of a predicate (*cover*) in context (*The article covered the coronavirus vaccine*) and automatically identify the frame it evokes (TOPIC). Though frame semantic parsing efforts have focused largely on the identification of semantic roles, frame identification is still an important task; it has been shown that a majority of errors in a complete frame semantic parsing system can be traced back to errors in frame identification (Hartmann et al., 2017).

In order to avoid the need for large scale annotation, we ask whether existing annotation from languages that are already well-covered (*supplementary languages*) can be re-used to train frame identification models in new languages (*target languages*). Recent multilingual embeddings are now providing a relatively simple technical means to seamlessly integrate training data from multiple languages (see Section 2.3 for details). However, it is much less clear whether the linguistic properties of the annotated datasets support this procedure. Often, FrameNet frames are found to be broadly applicable to other languages (Gilardi and Baker, 2018, Torrent et al., 2018); at the same time, some amount of 'tuning' may be required regarding their definition. To our knowledge, there are no studies that attempt to quantify these effects in models of cross-lingual frame identification. In linguistics, however, recent studies have emerged which present quantification of frame

transferability from English to Brazilian Portuguese on a preliminary study with a set of parallel, frame-annotated sentences (Torrent et al., 2018).

We operationalize the idea of quantifying transferability by training frame identification models on monolingual data (target language) and multilingual (target + supplementary languages), adopting a fixed *annotation budget* for the target language, where only a modest number of datapoints can be labeled. We fill the annotation budget by performing informed frame selection based on *frame transferability*. Our notion of frame transferability builds on work that estimates the difficulty of Word Sense Disambiguation (WSD) by measuring the coherence of the word senses in the data (McCarthy et al., 2016). Similarly, we assume that frames that are coherent in the supplementary languages will be better candidates for transfer to an unseen language, requiring less target language annotation than incoherent frames (see Section 2.4 for details).

Clearly, another prominent indicator of frame transferability would be a direct measurement of cross-lingual frame applicability (Boas, 2020, Sikos and Padó, 2018). Unfortunately, such methods already assume the existence of annotation in the target language. Therefore, we choose to exclude explicit measures of cross-lingual applicability from our models, since we crucially want our methodology to generalize to target languages for which we assume that no annotation is yet available. We later discuss cross-lingual frame comparability in our post-hoc analysis.

We select target annotations at the frame level (instead of selecting by predicates) for a few reasons. First, the frame level matches our goal of creating data to train a frame identification system. Second, in terms of data analysis, we are interested primarily in generalizable properties of frames rather than more fine-grained units.

In our empirical evaluation, we study frame identification over three target languages: English, German, and French, where the languages have frame definitions that are similar (taken directly from English) as well as different (adapted for the language of interest). Our selection method is based on latent properties of frame annotations, which reflect how the frame is used in context over each language. Therefore, we can evaluate which frames our selection models are more likely to choose for target language training: frames with similar or different definitions across languages.

Plan of the Paper. Section 2 sketches relevant related work. Section 3 contains the core method contribution of our study: A method for informed frame selection based on performance prediction using features for cross-lingual frame transferability. Section 4 describes the experimental setup, and Section 0 reports our results. Section 6 closes with a discussion.

2 Related Work

2.1 Frame-Semantic Analysis Across Languages

As sketched in the introduction, a prominent research question from a cross-lingual perspective is to what extent semantic frames can be considered to be 'universal' (Boas, 2020). Many FrameNet frames are found to be broadly applicable to other languages (Gilardi and Baker, 2018), and most projects considering other languages use some frames that are essentially unchanged from the English definition, alongside others that have been modified to suit the language of interest. Reasons that call for frame modifications include typological shifts or subtle differences in the

frame's interpretation which cause divergences in the core semantic roles and frame-evoking predicates (Ohara, 2014, Boas, 2005);

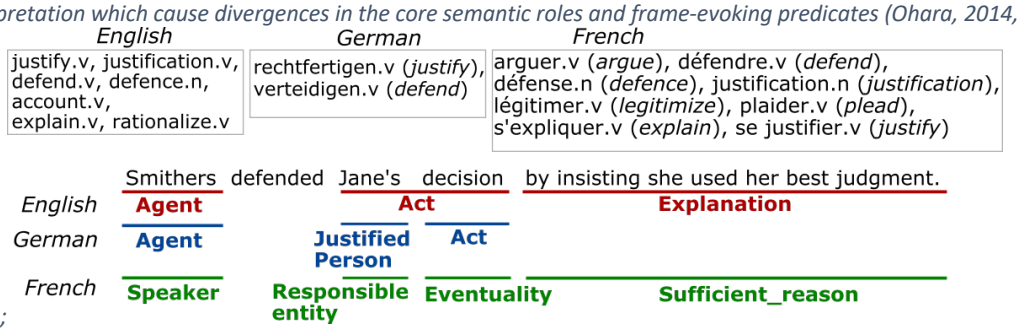


Figure 1 - JUSTIFYING frame in English, German, and French where the frame definition differs across all languages. The differences can be seen in terms of the frame-evoking predicates (above) and the core semantic roles (below).Figure 1 shows the JUSTIFYING frame in English (Baker, 2008), German (SALSA) (Burchardt et al., 2006), and French (Candito et al., 2014), where the definition has been modified for each language.

Differences in annotation strategies is another factor that affects the versatility and frequency of frame coverage in different frame semantic resources. Annotations typically proceed by a frame-by-frame approach, where the goal is decent coverage of each frame in the lexicon; lemma-by-lemma, where all senses of the annotated lemmas are covered; or full-text annotation, where frames are identified over running text. The English Berkeley FrameNet adopted both frame-by-frame and full text annotations, the French FrameNet used a frame-by-frame approach, and the German SALSA corpus took a lemma-by-lemma annotation approach.

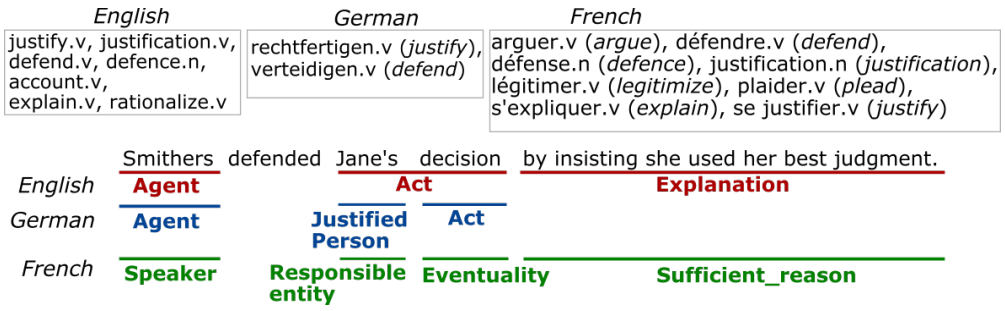


Figure 1 - JUSTIFYING frame in English, German, and French where the frame definition differs across all languages. The differences can be seen in terms of the frame-evoking predicates (above) and the core semantic roles (below).

2.2 Frame Semantics and Natural Language Processing

Frame semantics has been shown to benefit a number of downstream NLP tasks, including information extraction and question answering (Shen and Lapata, 2007, Burchardt et al., 2009b, Christensen et al., 2010, Taniguchi et al., 2018, Si and Roberts, 2018). Most recently, frames have been proposed as one of the frameworks that could be a basis for studying meaning construal, where the same conceptual background can be expressed with different emphasis or perspective (Trott et al., 2020).

To be useful at scale, though, all of these applications require accurate automatic

models of frame-semantic parsing, or at least frame identification. For the most part, all state-of-the-art models are based on word embeddings, high-dimensional representations of word meaning that are created from large collections of unstructured text. While previously such representations were directly based on counts, the current generation of word embeddings is based on neural network architectures such as Word2Vec (Mikolov et al., 2013), FastText (Bojanowski et al., 2017) or BERT (Devlin et al., 2019). Word embeddings can serve as input for supervised classification or regression models for specific tasks, whose training of course requires task-specific annotation ("fine tuning"). For frame identification, relatively straightforward embedding-based classification was quickly able to match and outperform traditional feature-based models (Hermann et al., 2014)[1].

Much of the recent work in frame identification focuses predominately on English, although resources have been developed in a handful of other languages – the largest and most well-covered include German (Burchardt et al., 2006), French (Candito et al., 2014), Dutch (Vossen et al., 2018), and Swedish (Borin et al., 2010). Following the release of these resources, frame semantic parsers were developed for most of these target languages, where classifiers predict frames with lexical and syntactic features (Johansson et al., 2012, Michalon et al., 2016, Erk and Padó, 2006).

2.3 Modeling Multilingual Frame Identification

The latest generation of embedding architectures are the so-called *transformers* which are able to learn contextual dependencies in an unsupervised fashion and construct context-dependent meaning representations: *tree* will receive one embedding in the phrase *the tree in the forest* and another one in the phrase *dependency tree*. Not surprisingly, one of the best-known transformer models, BERT (Devlin et al., 2019), is the basis of state-of-the-art frame identification models for English (Sikos and Padó, 2019, Tan and Na, 2019).

The simplest way to set up the BERT model for frame identification is to predict one frame (including a 'None' option) for each token in a sentence. In this setup, each training datapoint is a single annotated instance of a predicate and its context words, where the label that is predicted for the predicate is the correct frame. Such datapoints can be created straightforwardly from existing frame-semantic annotations.

An important recent development in word embeddings is *multilingual embeddings* (Upadhyay et al., 2016, Lample et al., 2018, Artetxe et al., 2020). Certain approaches to constructing multilingual embeddings involve adversarial training for refining embeddings cross-lingually (Lample et al., 2018), or bilingual dictionaries for transforming embeddings from a source to a target language (Artetxe et al., 2017). While BERT embeddings were initially trained on corpora in individual languages, researchers realized quickly that embeddings could be trained on multiple corpora simultaneously, or existing embedding spaces aligned with one another. In either case, the result is a space in which words from multiple languages are represented 'on par'. This enables the exploration of different scenarios including experiments where a model is trained with annotations from one

[1] Furthermore, the embedding approach generalizes to other modalities: (Botschen et al., 2018) use representations of images as a predicate's context to predict frames.

language and applied 'as-is' to another, so-called *zero-shot learning* (Wu and Dredze, 2019, Pires et al., 2019). For frame identification, this means that not even comparable corpora are necessary such as were used in previous approaches to cross-lingual frame identification (Johannsen et al., 2015, Kozhevnikov, 2016).

Recently, multilingual embeddings have been used to compute the alignment of lexical unit embeddings across languages in the Multilingual FrameNet alignment package[2]. These embeddings are based on large-scale, multilingual language models which we describe in our approach below, and the translation of a frame's lexical units across languages can be visualized by this method.

2.4 Frame Transferability

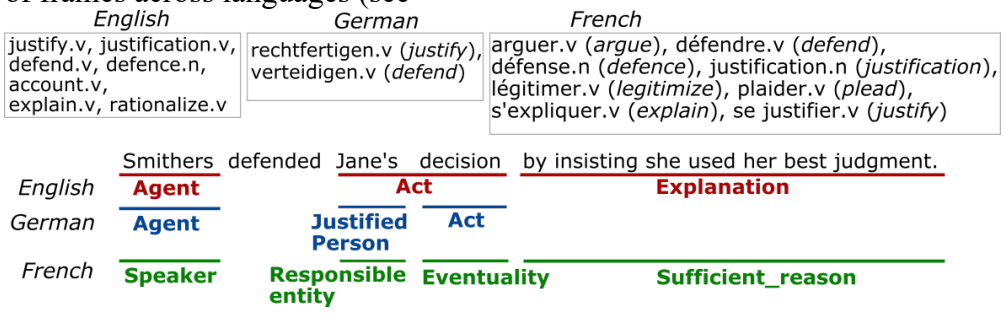
Since in frame identification, predicates can evoke multiple frames, this task bears a strong resemblance to the well-researched paradigm of WSD. This is why we use a study from WSD on the impact of semantic coherence on disambiguation difficulty (McCarthy et al., 2016) as our basis for estimating a frame's cross-lingual transferability in our multilingual frame identification models.

McCarthy et al. (2016) start from the observation that some words are much easier to disambiguate with regard to word sense than others. While factors like part-of-speech, frequency, or type of ambiguity (homonymy vs. polysemy) play a substantial role, a lot of variance remains unaccounted for. In response, they carry out a study in which they analyzed the difficulty of WSD for various lemmas in terms of the semantic coherence of the senses of these lemmas. They measured two aspects of coherence, representing senses as sets of embeddings for individual senses: (1), lemmas with senses whose instances form tight clusters should pose simpler WSD problems than lemmas whose senses are 'spread out'; and (2), lemmas whose senses are well separated from one another are presumably simpler to disambiguate. McCarthy et al. found very good empirical support for these hypotheses.

3 Methods

3.1 Cross-Lingual Frame Selection

Recall from Section 2.3 that our goal is to build a frame identification system for a *target language T*, while we assume that we have access to frame annotations for a set of *supplementary language(s) S*. The simplest way to do this would be to build a model using only the available frame-labeled data from *S*. However, given the imperfect comparability of frames across languages (see



[2] https://github.com/icsi-berkeley/framenet-multilingual-alignment

Figure 1), such a classifier will presumably not do well. Thus, our research question is: *Given a fixed annotation budget for T, how can we select frames for annotation to maximally improve a system that has only learned about frames from S*?

We pose our frame selection process as a performance prediction task (Bojar et al., 2017, Elloumi et al., 2018) where we are estimating a frame's cross-lingual transferability. We do this by estimating how much the annotations of a frame from *T* will improve frame identification given the availability of frame data from *S*. As such, frame selection is based on properties of the frames in *S* (which is the only data we assume we have), which we use to estimate how useful a *T* frame annotation will be towards improving the existing, multilingual frame identification system.

An overview of our approach is shown in Figure 2. It consists of three steps: building a baseline (we use the multilingual frame identification system from Section 2.3), learning the frame selection model where we estimate the transferability of frames and select frames from its estimations (Section 3.2), and using the selected *T* frames plus *S* frames to build a final frame identification model for *T*.

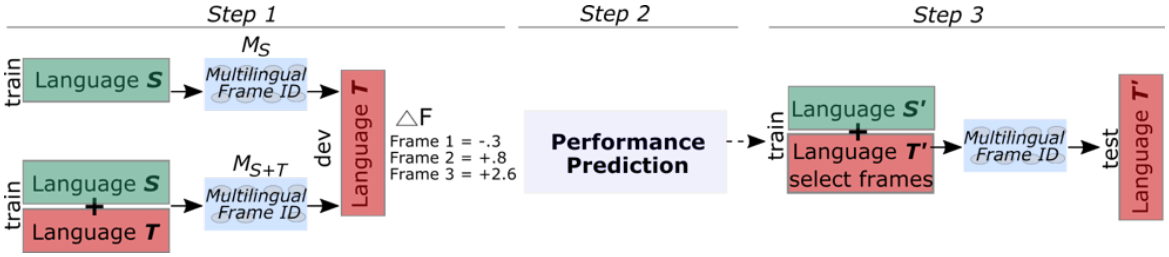


Figure 2 - Overview of Frame Selection

To learn the frame selection model, we need to use data from one language pair $\langle S,T \rangle$ for which we assume annotations already exist. We can then build multilingual frame identification systems trained (a) only on *S*, and (b) on *S* plus all available training data from *T*. We compare frame performance of these (a) and (b) systems to obtain ΔF, the change in performance by adding *T* frame annotations. A high Δ*F* indicates that the frame identification system benefits from the *T* annotations for that frame, whereas a low score indicates that the *S* annotations are already sufficient. Specifically, a high Δ*F* score suggests that the frame has a *lower* cross-lingual transferability, as more language-specific annotations are required to improve performance, and *S* annotations were not suitable for learning the frame.

In the general case, however, our goal is to define a frame selection process that generalizes to various target languages, including those for which no annotation is available at all. As we argued in the introduction, this means that we only use properties in the frame selection process that are based on data in the supplementary language *S*.

Finally, we can apply the frame selection model to rank the *T* frames by their estimated Δ*F* score and select the *T* frames with the highest scores for annotation. In our experiments in this article, we do not perform actual annotation; instead, we simulate annotation by simply sampling the respective frame annotations from the existing dataset. We then re-train a multilingual frame identifier on the *S* annotations, plus the annotation instances of the selected frames from *T*.

3.2 Estimation of Frame Transferability

Using the frame identification architecture described in Section 2.3, we train two models: one trained on all of *S* data, henceforth M_S, and a model trained on all of *S* plus the training set of *T* , henceforth M_S+T. We define ΔF for each frame *f* as the difference between the frame's F1 score from both models:

$$\Delta F = F1(f, M_S + T) - F1(f, M_S)$$

We compute ΔF of each frame over the development set in *T*[3]. A high ΔF indicates that a frame profits substantially from annotation in *T* and therefore has lower cross-lingual transferability.

Our frame selection is a linear regression model, which is a well-established architecture for data analysis in NLP (Baayen, 2008). Estimating frame transferability with linear regression also has the benefit of introspection into how frame properties are related to their performance.

3.2.1 Frame Transferability via Semantic Coherence

As introduced in Section 2.4, the properties that we consider are measures of semantic coherence following McCarthy et al. (2016). We replace the notion of 'sense' by the notion of 'frame', but use an analogous setup where each instance is represented by one (contextualized) embedding. Recall from Section 2.4 that McCarthy et al.'s first indicator was how tightly the instances of a word sense cluster together. Applied to frames, we have our first hypothesis concerning the variance of a supplementary language frame.

Hypothesis #1: *the larger the variance of a frame (i.e., the more dissimilar its instances to one another in the supplementary language), the more it profits from target language annotation*. We make this idea concrete as follows. Let centroid(*F*) be the average of all of its annotated instances *f*:

$$centroid(F) = \frac{1}{|F|} \sum_{f \in F} f$$

Then, we define *Var* as the variance of the frame by taking the difference between each individual frame instance (*f*) and its frame centroid:

$$Var(F) = \frac{1}{|F|} \sum_{f \in F} \|f - centroid(F)\|^2$$

The second indicator McCarthy et al. (2016) consider is the average of all between-cluster (i.e., between-sense) distances. We believe that for frame identification, where typically a small number of senses are realistic candidates, it is more sufficient to consider the separability between the current frame and its nearest neighbor. Therefore, we next hypothesize that distance affects frame performance. *Hypothesis #2*: *the smaller the distance*

[3] We use F1 scores to compute ΔF because we do not want frame selection to be biased by frames that are highly frequent in the target language test data, and accuracy would be ill-suited because of the dominant number of true negatives. For final evaluation, however, we still use the established measure for frame identification, which is overall accuracy over instances (see Section 4.4).

between a frame and its nearest neighbor, the more it profits from target language annotation. Formally, we define *Dist* as the distance between frame centroids, calculated by cosine similarity between a frame *F* and its nearest neighbor *F′*:

$$Dist(F) = \|centroid(F) - centroid(F')\|^2$$

As a third indicator, we compute the coherence of a frame as the ratio of the *Dist* and *Var* scores:

$$Co(F) = {Dist(F)}/{Var(F)}$$

We include *Co* to account for interactions between *Dist* and *Var* and again assume a larger benefit of target language annotation for lower values of *Co*. Concretely, if we assume in Hypothesis 1 that variance of supplementary language frames should be high, and Hypothesis 2 that distance of frames and their nearest neighbors should be low, we would predict that frames with the lower *Co* values would be better candidates for frame selection. Alternatively, a higher value of *Co* would indicate that the frame already has good clusterability, with low variance across the frame's instances and a high distance from other frames, and therefore would likely be learnable from the supplementary annotations and wouldn't require additional target language data.

4 Experimental Setup

4.1 Experimental Rationale

As we described in Section 3.1, we start with only frame annotations from *S* and subsequently add a moderate budget of annotations from *T* (we consider budget sizes of 5k and 10k instances). We simulate target language "annotation" by taking randomly sampled annotated instances of each selected frame from *T*. In certain cases, there can be a high number of annotations for a single frame; in fact, some resources have frames with a very high (>1000) number of annotations. If we take all the training instances from these frames, we reduce the diversity of frames that are seen by the classifier and the added frame data would be dominated by these few, highly annotated frames. To prevent this problem, we restrict the number of instances of each frame to 200 random instances, motivated by a desire to cover a substantial number of frames. The number 200 was selected to balance the goals of adding a substantial number of frames and a substantial number of instances per frame.

Since our experiment uses informed frame selection, the question remains how we train the frame selection model. As we noted in Section 3.1, frame identification training requires annotated data both for *S* (to provide the features) and *T* (to evaluate the predictions). We therefore train the frame selection model on our language pair ⟨*S, T*⟩ with the largest number of overlapping frames, namely ⟨German, English⟩. We use the development set of *T* (in this case, English) to learn the frame selection model so that there is no information leakage to either frame identification model training or frame identification model evaluation. The frame selection model is then applied as-is to all other language pairs ⟨*S′,T′*⟩ for frame

selection, thus demonstrating its generalization capabilities to unseen languages. Models are then trained with this modified $\langle S'+selected\ T' \rangle$ data and evaluated over unseen T' test data.

Below, we present results for all combinations of supplementary and target languages. Due to our use of multilingual embeddings, we can also construct models based on multiple supplementary languages for a single target language. For these models, we combine the ranked list of frames from each individual $\langle S', T' \rangle$ pairs and take the top predicted frames from this combined set as our selected frames.

4.2 Datasets

Despite growing efforts to create frame-semantic resources for different languages (Torrent et al., 2018), the number of languages with sufficient amounts of publicly available frame semantic annotations suitable for NLP models is still limited. For this reason, our experiments cannot rival the massively multilingual setups that have been explored for word embeddings (Ammar et al., 2016) or parsing (Agić et al., 2016). Another practical limitation we encountered was the familiarity of the authors with the languages under consideration to qualitatively assess and analyze the output of the models. Therefore, we focus on frame-semantic annotations for three languages: the Berkeley FrameNet 1.5 annotations for English[4], the French FrameNet corpus[5] (Candito et al., 2014) and the German SALSA corpus[6] (Burchardt et al., 2006).

Table 1 provides descriptive statistics for the three resources, including numbers for the frame overlap with English.

	#	# instances			Frame overlap w/ English		
T Lang	frames	Train	Dev	Test	same	mod	unaligned
EN	1020	15044	4434	4458	-	-	
FR	105	16961	1732	2941	46	22	37
DE	1001*	26070	5530	5659	256	37	730

*Table 1 - Frame-semantic resources for English (EN), French(FR), German(DE) where frames that have not been modified from the English definitions ("same") and frames that have been modified ("mod") represent large subsets of the frames in the FR and DE resources. Language specific frames in are not aligned cross-lingually ("unaligned"). *Total for frames with "same/mod/unaligned" is higher than the # frames, as there are frames in both "same/mod" categories in DE (discussed further in Section 5.3.1).*

4.2.1 Berkeley FrameNet 1.5

The FrameNet 1.5 full-text annotations form the standard corpus for frame identification systems in English and cover a bit more than 1000 frames. In our training, we use a single frame-evoking element, its sentential context, and its frame as one instance for the classifier. We adopt the widely used test/train/dev splits defined by Das et al. (2014).

4.2.2 French FrameNet

The French FrameNet project (Djemaa et al., 2016) adapted their frame inventory from the English FrameNet 1.5. Frame annotations were added to the French Treebank and

[4] https://framenet.icsi.berkeley.edu/

[5] https://sites.google.com/site/anrasfalda/

[6] http://www.coli.uni-saarland.de/projects/salsa/corpus/request/salsa-corpus-request.cgi

Sequoia treebank (Abeillé and Barrier, 2004, Candito and Seddah, 2012), which covered four domains (commercial transactions, cognitive stances, causality, and verbal communication). French FrameNet provides its own test/train/dev splits.

The French data covers only about 100 frames annotated compared to roughly 1000 frames for the two other languages, resulting from a different sampling strategy. Many French frames were adopted as-is from the Berkeley FrameNet, but about half of them were systematically restructured to yield a better fit with the corpus. This includes cases where multiple English frames have been combined into a new frame. For example, French has the CHATTING_DISCUSSION frame, which combines the CHATTING and DISCUSSION frames from the English lexicon; such cases count as 'unaligned' in the table. Since the number of annotated instances for French and German is on the same order of magnitude (within a factor of 2), but the number of French frames is substantially lower, the average number of annotated instances per frame is highest for French. We believe that this combination of properties (close to English but many changed frames) makes French an interesting target language in our experiments.

4.2.3 The German SALSA corpus

The SALSA corpus provides frame-semantic annotations over the German TIGER news corpus (Brants et al., 2002). We use the train/test/dev splits defined by Botschen et al. (2018) for our experiments.

SALSA initially adopted frames from the English FrameNet 1.2 inventory. A comparatively small number of frames was modified; in contrast, a large number of frame approximations, called "proto-frames", was added (these count as 'unaligned'). These are lemma-specific frame structures developed to cover instances for which FrameNet did not provide an adequate frame (see Burchardt et al. 2009a for details).

4.3 Multilingual Embeddings

As embeddings, we use mBERT, a multilingual BERT model which represents words of over 100 languages in a shared semantic space. This model was trained on Wikipedia dumps available for the various languages (Karthikeyan et al., 2019).

4.4 Evaluation

Classifier accuracy is the percentage of correct predictions of the classifier when the full set of classes is used, and is a standard metric of evaluation for computational systems. For frame identification, we use the full set of frame classes, meaning there is no assumption about which specific frame candidates a single predicate might evoke.

4.4.1 Baselines

We report several different baselines for our experiment. The *S only* baseline only uses data from the supplementary *S* language(s) and tests on a target *T* language without any *T* training data. Frames that are used from the *S* language for training in *T* are the frames in *S* data that are shared with the target language. These include the "same" and "modified" frames in Table 1, where we do not include frames that are "unaligned" in French and German. For German, the "unaligned" cases include language-specific, "proto-frames", and for French, these include frames whose definition is a blend between two frames where the frame is

essentially language-specific and not readily alignable to an English or German frame[7]. In other words, frames that can be readily mapped back to a *T* frame through the frame's naming and semantic roles are used for *S only* training.

The *Random* baseline adds 5k or 10k instances of randomly selected *T* frames to the *S only* data. Identical to the *Embedding* model, a maximum of 200 random instances per frame are chosen.

[7] This work was conducted before the release of the Multilingual FrameNet alignment tool, which for future extensions, could be an additional measure for cross-lingual frame alignment

5 Results

			S only		*S+T with Frame Selection*			
					Random		*Embeddings*	
	T	*S*	#	*all*	+5k	+10k	+5k	+10k
1	EN	DE	25k	17.88	33.99	55.61	**35.76**	**60.77**
2		FR	11k	3.99	27.74	52.13	**28.29**	**57.94**
3		DE+FR	36k	14.27	30.80	59.14	**54.40**	**62.36**
4	DE	EN	14k	38.79	23.33	37.06	**24.16**	**42.75**
5		FR	8k	5.57	18.45	35.27	**21.66**	**40.25**
6		EN+FR	22k	27.99	22.55	39.54	**43.24**	**43.88**
7	FR	DE	7k	12.42	37.88	54.97	**47.03**	**62.09**
8		EN	2k	17.38	25.58	59.26	**46.65**	**59.66**
9		DE+EN	9k	18.55	25.76	59.25	**59.99**	**61.77**
	a	*b*	*c*	*d*	*e*	*f*	*g*	*h*

Table 2- Results for frame selection(baselines and cross-lingual training): Test set classifier accuracies for models using all supplementary data (S only) with number of S instances used in training (#) and supplementary data plus a fixed budget (5k/10k) of target annotations (S+T) selected by different criteria (random, embedding-based features).

The starting point of our experiment is the baseline which used only training data of the supplementary language (*S only*) and evaluated on the test data of the target language (*T*). Results for this setting are given in the *all* column in Table 2 (cells 1d – 9d). For each target (*T*) language, *S only* results are given for all supplementary languages, including combined supplementary languages. In all *T* languages, we see that learning with the supplementary annotations alone achieves accuracies of 4% from FR to EN (where FR is *S* and EN is *T*, cell 2d) to 39% from EN to DE (cell 4d). The comparably bad results for FR as *S* are mainly a result of the small frame intersections between FR and the other languages (cf. **Error! Reference source not found.**). Conversely, EN to DE has the largest frame intersection, and DE is the best model for English as the target language, presumably due to this higher number of shared frames. In sum, leveraging annotations from different, supplementary languages alone - that is, assuming that no annotations for the target language are available, shows reasonable performance but arguably does not yield models that are practically usable.

We therefore proceed with adding target language annotations (+5k and +10k) back to the multilingual training. We first consider random frame selection to disentangle the effect of added *T* data in general with the effect of a deliberate selection of frames (cells 1e – 9f). Without comparing our selection to a random frame selection, it would remain an open question as to whether no selection of frames was necessary in the first place and that any target language data of a certain size would yield comparable improvements. Compared to the *S only* training, results in *Random* show that even with a random selection of 5k instances from *T* the performance achieves significant gains. However, all language pairs benefit from a more informed frame selection (cells 1g – 9h). Regarding the effect of dataset size, we

unsurprisingly find that adding more data (+10k) is always better than adding fewer data (+5k) within each selection strategy, although the improvement is smaller in cases where data from multiple languages is combined (DE+EN to FR and EN+FR to DE). However, in those cases, selecting +5k instances ranked by the embedding-based predictors actually yields a higher accuracy than +10k instances from random frames.

In terms of language pairs, we observe that results for EN to DE and DE to EN are consistently higher than for FR to DE and FR to EN, respectively. When using French as the target language, none of the two supplementary languages performs consistently better than the other. In combination, however, we observe the highest improvements for French. In general, the best results for each language *T* combine both *S* languages. This suggests that, when only a few annotations are available, a new language would likely benefit the most from a simple concatenation of available frame annotations from various source languages. In fact, a modest +5k instances of data from multiple languages could achieve similar results as +10k instances depending on the source and target languages.

5.1 Benefit of Supplementary Language

Our approach uses the supplementary annotations in two capacities: 1) as part of the multilingual training data, and 2) for the selection of informative target frames. One unanswered question from the results presented in Table 2 is whether the supplementary data is actually benefiting the system at all; more specifically, we need to ask whether we would have achieved the same results with a selection of *T* frames alone. To answer this question, we train *T only* models which train the classifiers only on the same 5k/10k instances used for *Random* baseline, without using any supplementary data. The results of these tests are given in Table 3 below and are directly comparable to the Random baseline results in Table 2 (repeated below for clarity). For the S+T setting from Table 2, we show performance of the combination of both supplementary languages for each target language (cells 3a,f/6a,f/9a,f).

Training data	*Target Language*					
	EN		*DE*		*FR*	
T only (random frame selection)	+5k	+10k	+5k	+10k	+5k	+10k
	24.92	47.75	18.64	34.56	25.12	48.05
S+T (random frame selection)	+5k	+10k	+5k	+10k	+5k	+10k
	30.80	59.14	22.55	39.54	25.76	59.25

Table 3 - T only model results: test set classifier accuracy when training only on target language (T only), compared to the best performing Random baseline of selected frames (Random baseline).

Table 3 shows performance in the *T only* training is consistently and, in most cases substantially, lower than performance for the target languages when S data is added (S+T). This demonstrates again the benefits of multilingual training and confirms that it is worth using multilingual data for training frame identification models when it is available.

5.2 Analysis of the Frame Selection Model

Finally, we ask whether we can analyze the performance prediction model in order to better understand how embedding properties of frames are related to the improvement for this frame when adding target language annotation, Δ*F*. Unfortunately, it turns out that the three

properties that we have defined (coherence, nearest neighbor distance, and within-frame variance) show a high degree of collinearity – which is not surprising, given that coherence is defined as a ratio of the other two properties. As a consequence, the coefficients of the performance prediction model lose their interpretability (e.g., McNamee 2005).
For this reason, we excluded the coherence of a frame (*Co(F)*) from this analysis and estimated a simpler model including only two normalized predictors, namely nearest neighbor distance (*Dist*) and within-frame variance (*Var*). The results are shown in Table 4. We initially hypothesized (cf. Section 3.2.1) that 1) the more dissimilar the instances of the frame are to one another, the more it will profit from target language annotation, and 2) the smaller the distance between a frame and its nearest neighbor, the more it will profit from target language annotation. The coefficients confirm only Hypothesis #1, where a high within-frame variance is very significant in predicting a higher ΔF. The other property (*Dist*) does not significantly contribute to the prediction of ΔF, indicating that the separation from the nearest neighbor frame is possibly an oversimplification as a measure of the difficulty to model a frame.

Predictor	*Coeff*	*Std. Error*	*p value*
Nearest neighbor distance (*Dist*)	0.005	0.07	>0.10
Within-frame variance (*Var*)	0.21	0.07	<0.01

Table 4 - Estimated coefficients and p-values for two embedding-based frame properties in a simplified performance prediction linear regression model.

5.3 Analysis of Frame-Level Performance

We now proceed with an analysis of the frame transfer method and the comparability of frames at the lexicographic level – that is, how well frame definitions are aligned across languages. While the transfer method relied solely on available annotations in the supplementary language, our analysis below looks at the lexicon in both languages, where we compare the performance of frames with high cross-lingual comparability in terms of their lexicographic entries versus frames that are thought to have low lexicographic comparability.

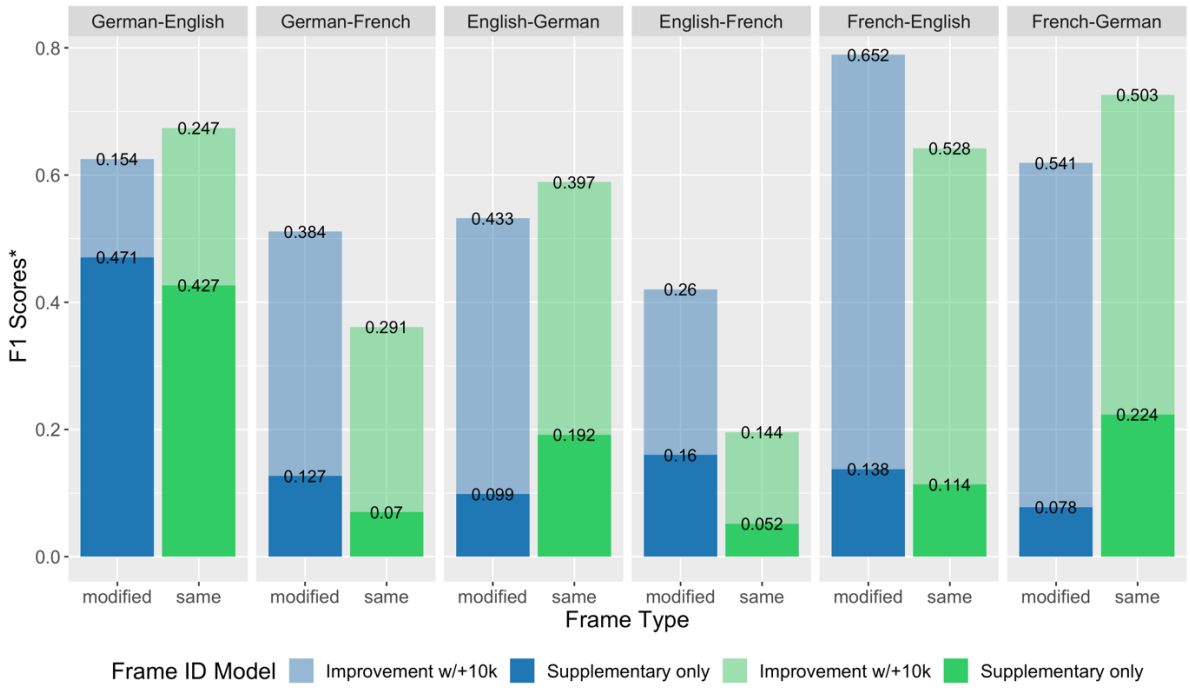


Figure 3 - Correlation between similarities in the frame definitions ("Frame Type") and model performance ("F1 Scores") for frames selected for annotation. Frames are either modified across languages and therefore diverge lexicographically ("modified") or they have the same definition across language pairs ("same"). Language pairs are in the form <T-S> (e.g., DE-EN is DE as T and EN as S) where results are tested over T test data. For the "Supplementary only" condition (dark bars), we report absolute F1 scores for performance, while "Improvement w/+10k" shows the average increase in F1 score (light bars) after the frame type was added.

5.3.1 Similarity in Frame Definitions

The German and French FrameNets distinguish between frames that have been modified from the original English FrameNet definition and those that are consistent with English. We take the frames that were selected for annotation in the target language and ask whether there is a difference in the performance gains across these two frame types ("same" and "modified" in their cross-lingual definition). In Table 1, there are 22 cases of German frames that are listed in both categories; for example, the COGITATION frame has two entries in SALSA, one with modified semantic roles and the other which has retained the English definition. We disregard these cases from our analyses.

Figure 3 shows that, for all language pairs except one (EN-DE), the selection of modified frames led to higher improvement. The JUSTIFYING frame, where the definition diverges across all three languages (showed in Figure 1), is one of the frames consistently selected by our model for all language pairs. One possible reason for this is that the frames which are described as the same across the resources are already learned sufficiently by *S*, leading to lower gains in multilingual training; for instance, the CAUSATION frame was not modified across any language pair, and was never selected as a target for further, language-specific annotations. When we compare absolute F1 scores of the *S only* model, the results are mixed – only two of the language pairs support this hypothesis (FR-DE, EN-DE), while other language pairs (FR-EN, DE-EN) show similar F1 scores for both frame types. However, modified frames predominately benefit from the target language

annotations, suggesting that researchers building frame semantic resources for different languages should focus more on these modified frames. If it is the case that researchers should target modified frames for annotation, the question might then arise: how would they know whether a frame should be modified?

Evidence from previous studies suggest that typological differences between languages can be expected to affect the frame lexicon in a target language (Boas, 2005, 2020), but those typological differences can be predictable to a certain extent. Hasegawa et al. (2011) identify cases of frames in English that are primarily composed of transitive verbs and tend to translate poorly in Japanese because Japanese typically prefers to describe events as stative (Ikegami, 1991). These frames would be expected to require modification if one were to build a frame lexicon in Japanese. Beyond typological differences, analysis of parallel corpora has indicated substantial freedom for translators regarding the linguistic realization of the same event: Torrent et al. (2018) find shifts in the part of speech of certain frame-evoking lemmas to cause different frame assignments across translations; Padó and Erk (2005) investigates cases where the contribution of a single frame-evoking element is split among multiple frame-evoking elements in translation. Systematic mining of parallel and comparable corpora could make it feasible for researchers working on a target language to get an idea of specific frames that could require modification, and therefore would warrant annotation.

5.3.2 Frames with high/low performance in S only training

We take results from the *S only* model to see which frames performed best across different language pairs. In this condition, no *T* annotations were used in training, but frame performance is measured over *T*. As shown in Table 5, many of the frames with the highest F1 scores across the EN-DE pair are those whose predicates form a tight semantic cluster; for example, the KINSHIP frame whose predicates are all familial relationships (*brother, sister, grandfather,* etc.) or the PEOPLE frame which consists of terms relating to humans (*man, woman, child*). Frames that perform well with only supplementary data are those with low variance within a frame (tight clustering of its instances - in this case, predicates), indicating that they are easier to learn when they form a tight cluster. This is opposite to the results we find in the performance prediction model, where we predict the frames with high variance will need more target language data to learn. Other explanations of these results include the fact that the lexical units in these frames are largely nominal, and their valency patterns are less likely to differ significantly across languages.

Performance for French frames are harder to interpret. Recall from Section 4.2.2 that the set of annotated frames in French was limited to four specific domains. Many of the high performing French frames (COMMERCIAL_TRANSACTION, COMMERCE_BUY, COMMERCE_SELL, IMPORTING) are in the commerce domain, while frames from cognitive stances or communication (QUESTIONING, REGARD, COMMUNICATION_RESPONSE, JUDGMENT_DIRECT_ADDRESS, CONTACTING) tend to appear as low performing cross-lingually. However, the change in domain covaries with other properties: The majority of lexical units (60%) from the commerce domain are nominal predicates, while predicates from the cognitive stance and communication domains are largely verbal (only 28% and 23% nominal, respectively) (Djemaa et al., 2016). This

aligns with observations from EN-DE, where the part of speech of the lexical units across languages has a strong impact on cross-lingual performance. It is also possible that the predicates are clustered more tightly in the commerce domain than the other three domains. Ultimately, however, the small number of French frames does not admit a strong interpretation of these findings.

German					
EN	*High*	*Low*	*FR*	*High*	*Low*
	MEMBERSHIP PEOPLE CALENDRIC_UNIT	TOPIC EXPERTISE FILLING		COMMUNICATION_RESPONSE TEXT_CREATION REASON	COMING_TO_BELIEVE QUESTIONING JUDGMENT_DIRECT_ADDRESS
English					
DE	*High*	*Low*	FR	*High*	*Low*
	PART_WHOLE PEOPLE_BY_AGE KINSHIP	TAKING_TIME SIMILARITY JUSTIFYING		EXPORTING COMMERCIAL_TRANSACTION ATTRIBUTED_INFORMATION	ENCODING DESERVING REGARD
French					
EN	*High*	*Low*	*DE*	*High*	*Low*
	COMMERCIAL_TRANSACTION DECIDING IMPORTING	CONTACTING PROVING CAUSE_EARNING		COMMERCE_BUY COMMERCE_SELL REFERRING_BY_NAME	JUSTIFYING COMING_TO_BELIEVE COMMUNICATION_RESPONSE

Table 5 - Frames with top F1 scores from the S only model (High) and the lowest F1 scores (Low). Columns (EN, FR, DE) show S languages, rows (German, English, French) show the target (T) languages.

6 Conclusion

The question of the universality of frames has been posed since the beginning of the theory of frame semantics (Fillmore, 1982, Boas, 2005, 2020). In fact, comparable frames have been found across even typologically unrelated languages such as English and Japanese, presumably due to the fact that frames allow a certain degree of variation in how they can be expressed (Hasegawa et al., 2014). At the same time, frame identification and, more broadly, frame semantic parsing, all require annotated data. Many languages do not have the resources to invest in a full-scale frame annotation project that would lead to a practically usable automatic frame identification system. As computational linguists, we can ask whether we can supplement some of the annotation needs for a target language by existing annotations in other languages.

This study was, to our knowledge, the first one to investigate this question of learning frame identification models based on multilingual embeddings. We defined a method that selects frames for annotation in the target language based on estimates of a frame's

transferability. To make this estimate, we use features of semantic coherence. Compared to a setting in which we do not use any target language annotation (which yields promising but still ultimately low performance), we found that informed frame selection can construct usable frame identification models within a manageable annotation budget. The most important factor in frame selection, according to our model, is frame-internal variance: Frames which have a more compact cluster in the supplementary language, meaning their predicates all form a relatively coherent group, require less target language annotation than frames that were spread out more. We find this is the case even when the number of instances per frame (200) that we randomly select is relatively modest, and the number of frames (25 frames maximum in 5k, 50 frames maximum in 10k) are also modest. This validates our approach that one can still see improvement in target language frame identification with only a modest, fixed number of frame instances.

In a post-hoc analysis, we established that, overwhelmingly, the frames that were selected yield better results when they have lexicographic definitions which diverge across languages. One plausible explanation for this result is that these lexicographic modifications were motivated by typological differences across the language pairs such as lexicalization or syntactic valence, which emerge as divergences in the semantic representations of the frames in the computational model. Therefore, these modified frames are more useful for selection, as they help refine a supplementary-based language model to learn the specific properties of frames for the target language.

It cannot be overlooked that the makeup of the frame annotations themselves could have played a large role in the utility of cross-lingual data for frame identification. While much prior work in computational linguistics has shown that datasets with sometimes significant divergences in certain semantic role labeling schema (a subtask of frame semantic parsing) can still be combined for improved results (Akbik and Li, 2016, Feizabadi and Padó, 2015), we find that the combination of different frame annotations alone does not lead to the greatest possible gains. In fact, there are significant differences in the numbers of instances of each frame that have been annotated, as well as the variety of predicates that evoke those frames. For instance, the German SALSA resource (Burchardt et al., 2006) has one frame (POLITICAL_LOCALES) with nearly 1k annotations for a single predicate (*Land.n*), while each predicate is annotated exactly 100 times in the French FrameNet (Candito et al., 2014). While we controlled for these differences in our selection method by only taking a random sample of 200 instances per frame, it is possible that these differences have an effect when only using supplementary language annotations. Future work could involve controlling for these effects by taking only a fixed number of frame instances from supplementary data in training for a target language.

Our study considered three languages that are among those languages with the largest frame-semantic resources (English, German, and French). It is clear that generalization of our results must consider that these languages are typologically close to one another (although see Burchardt et al., 2009a), and many potential target languages are more dissimilar to these supplementary languages. Naturally, an important avenue of future research is the generalization of our frame selection to a broader range of target languages. As we described earlier, the Multilingual FrameNet alignment tool (described in Section 4.4.1) could be another promising way to gauge frames that would require more annotation for target language frame identification, as these frames would have poorer cross-lingual

alignment of their lexical units. However, it would be straightforward to extend our framework to other languages as we observe that a target language model already sees impressive gains with 5k instances of annotated data, which is a small requirement for frame annotation.

References

Abeillé, Anne and Nicolas Barrier. 2004. Enriching a French treebank. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*. Lisbon, Portugal: European Language Resources Association (ELRA).

Agić, Željko, Anders Johannsen, Barbara Plank, Héctor Martínez Alonso, Natalie Schluter, and Anders Søgaard. 2016. Multilingual projection for parsing truly low-resource languages. *Transactions of the Association for Computational Linguistics* 4:301–312.

Akbik, Alan and Yunyao Li. 2016. POLYGLOT: Multilingual semantic role labeling with unified labels. In *Proceedings of ACL-2016 System Demonstrations*, pages 1–6. Berlin, Germany: Association for Computational Linguistics.

Ammar, Waleed, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A. Smith. 2016. Massively multilingual word embeddings. *CoRR* abs/1602.01925.

Artetxe, M., Labaka, G., & Agirre, E. (2017, July). Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 451-462).

Artetxe, M., Ruder, S., & Yogatama, D. (2020, July). On the Cross-lingual Transferability of Monolingual Representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 4623-4637)

Baayen, R. H. 2008. *Analyzing Linguistic Data: A Practical Introduction to Statistics using R*. Cambridge University Press.

Baker, Collin. 2008. FrameNet, present and future. In J. Webster, N. Ide, and A. C. Fang, eds., *The First International Conference on Global Interoperability for Language Resources*. City University, Hong Kong: City University.

Ben Aharon, Roni, Idan Szpektor, and Ido Dagan. 2010. Generating entailment rules from FrameNet. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 241–246. Uppsala, Sweden: Association for Computational Linguistics.

Boas, Hans C. 2005. Semantic Frames as Interlingual Representations for Multilingual Lexical Databases. *International Journal of Lexicography* 18(4):445–478.

Bojar, Ondřej, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Shujian Huang, Matthias Huck, Philipp Koehn, Qun Liu, Varvara Logacheva, Christof Monz, Matteo Negri, Matt Post, Raphael Rubino, Lucia Specia, and Marco Turchi. 2017. Findings of the 2017 Conference on Machine Translation. In *Second Conference on Machine Translation*, pages 169–214. The Association for Computational Linguistics.

Borin, Lars, Dana Dannélls, Markus Forsberg, Maria Toporowska Gronostaj, and Dimitrios Kokkinakis. 2010. The past meets the present in swedish FrameNet++. In *14th EURALEX international congress*, pages 269–281.

Botschen, Teresa, Iryna Gurevych, Jan-Christoph Klie, Hatem Mousselly-Sergieh, and Stefan Roth. 2018. Multimodal frame identification with multilingual evaluation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1481–1491. New Orleans, Louisiana: Association for Computational Linguistics.

Brants, Sabine, Stefanie Dipper, Silvia Hansen, Wolfgang Lezius, and George Smith. 2002. The tiger treebank. In *Proceedings of the workshop on treebanks and linguistic theories*, vol. 168.

Burchardt, Aljoscha, Katrin Erk, Anette Frank, Andrea Kowalski, Sebastian Padó, and Manfred Pinkal. 2006. The SALSA corpus: a German corpus resource for lexical semantics. In *Proceedings of LREC*, pages 969–974.

Burchardt, Aljoscha, Katrin Erk, Anette Frank, Andrea Kowalski, Sebastian Padó, and Manfred Pinkal. 2009a. Using FrameNet for the semantic analysis of German: Annotation, representation, and automation. In H. C. Boas, ed., *Multilingual FrameNets in Computational Lexicography – Methods and Applications*, pages 209–244. De Gruyter.

Burchardt, Aljoscha, Marco Pennacchiotti, Stefan Thater, and Manfred Pinkal. 2009b. Assessing the impact of frame semantics on textual entailment. *Natural Language Engineering* 15(4):527–550.

Candito, Marie, Pascal Amsili, Lucie Barque, Farah Benamara, Gaël de Chalendar, Marianne Djemaa, Pauline Haas, Richard Huyghe, Yvette Yannick Mathieu, Philippe Muller, Benoît Sagot, and Laure Vieu. 2014. Developing a French FrameNet: Methodology and first results. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 1372–1379. Reykjavik, Iceland: European Language Resources Association (ELRA).

Candito, Marie and Djamé Seddah. 2012. Le corpus sequoia: annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical (the sequoia corpus: Syntactic annotation and use for a parser lexical domain adaptation method) [in French]. In *Proceedings of the Joint Conference JEP-TALN-RECITAL 2012, volume 2: TALN*, pages 321–334. Grenoble, France: ATALA/AFCP.

Christensen, Janara, Stephen Soderland, Oren Etzioni, et al. 2010. Semantic role labeling for open information extraction. In *Proceedings of the NAACL HLT 2010 First International Workshop on Formalisms and Methodology for Learning by Reading*, pages 52–60. Association for Computational Linguistics.

Das, Dipanjan, Desai Chen, André F. T. Martins, Nathan Schneider, and Noah A. Smith. 2014. Frame-semantic parsing. *Computational Linguistics* 40(1):9–56.

Devlin, Jacob, 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.

Djemaa, Marianne, Marie Candito, Philippe Muller, and Laure Vieu. 2016. Corpus annotation within the French FrameNet: a domain-by-domain methodology. In *Tenth International Conference on Language Resources and Evaluation (LREC 2016)*. Portorož, Slovenia.

Elloumi, Zied, Laurent Besacier, Olivier Galibert, Juliette Kahn, and Benjamin Lecouteux. 2018. Asr performance prediction on unseen broadcast programs using convolutional neural networks. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5894–5898. IEEE.

Ellsworth, Michael and Adam Janin. 2007. Mutaphrase: Paraphrasing with FrameNet. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 143–150. Prague: Association for Computational Linguistics.

Erk, Katrin and Sebastian Padó. 2006. Shalmaneser-a toolchain for shallow semantic parsing. In *LREC*, pages 527–532.

Feizabadi, Parvin Sadat and Sebastian Padó. 2015. Combining seemingly incompatible corpora for implicit semantic role labeling. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 40–50.

Fillmore, Charles J. 1982. *Frame semantics*, pages 111–137. Seoul, South Korea: Hanshin Publishing Co.

Fillmore, Charles J. and Collin F. Baker. 2001. Frame semantics for text understanding. In *Proceedings of the NAACL WordNet and Other Lexical Resources: Applications, Extensions and Customizations Workshop*.

Gilardi, Luca and C Baker. 2018. Learning to align across languages: Toward multilingual FrameNet. In *Proceedings of the International FrameNet Workshop*, pages 13–22.

Gruzitis, Normunds, Gunta Nespore-Berzkalne, and Baiba Saulite. 2018. Creation of Latvian FrameNet based on universal dependencies. In *Proceedings of the International FrameNet Workshop (IFNW)*, pages 23–27.

Hasegawa, Yoko, Russell Lee-Goldman, and Charles J Fillmore. 2014. On the universality of frames: evidence from english-to-japanese translation. *Constructions and Frames* 6(2):170–201.

Hermann, Karl Moritz, Dipanjan Das, Jason Weston, and Kuzman Ganchev. 2014. Semantic frame identification with distributed word representations. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1448–1458. Baltimore, Maryland: Association for Computational Linguistics.

Johannsen, Anders, Héctor Martínez Alonso, and Anders Søgaard. 2015. Any-language frame-semantic parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2062– 2066. Lisbon, Portugal: Association for Computational Linguistics.

Johansson, Richard, Karin Friberg Heppin, and Dimitrios Kokkinakis. 2012. Semantic role labeling with the swedish framenet. In *LREC*, pages 3697– 3700.

Karthikeyan, Kaliyaperumal, Zihan Wang, Stephen Mayhew, and Dan Roth. 2019. Cross-lingual ability of multilingual bert: An empirical study. In *International Conference on Learning Representations*.

Lample, G., Conneau, A., Denoyer, L., & Ranzato, M. A. (2018, February). Unsupervised Machine Translation Using Monolingual Corpora Only. In *International Conference on Learning Representations.*

McCarthy, Diana, Marianna Apidianaki, and Katrin Erk. 2016. Word sense clustering and clusterability. *Computational Linguistics* 42(2):245–275.

McNamee, Roseanne. 2005. Regression modelling and other methods to control confounding. *Occupational and Environmental Medicine* 62(7):500–506.

Michalon, Olivier, Corentin Ribeyre, Marie Candito, and Alexis Nasr. 2016. Deeper syntax for better semantic parsing. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 409–420. Osaka, Japan.

Ohara, Kyoko. 2014. Relating frames and constructions in japanese framenet. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2474–2477.

Padó, Sebastian and Katrin Erk. 2005. To cause or not to cause: Cross-lingual semantic matching for paraphrase modelling. In: *Proceedings of the Cross-Language Knowledge Induction Workshop*. Cluj-Napoca, Romania.

Pires, Telmo, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001. Florence, Italy: Association for Computational Linguistics.

Roth, Michael and Mirella Lapata. 2015. Context-aware frame-semantic role labeling. *Transactions of the Association for Computational Linguistics* 3:449–460.

Shen, Dan and Mirella Lapata. 2007. Using semantic roles to improve question answering. In *Proceedings of the 2007 joint conference on empirical methods in natural language processing and computational natural language learning (EMNLP-CoNLL)*, pages 12–21.

Si, Yuqi and Kirk Roberts. 2018. A frame-based nlp system for cancer-related information extraction. In *AMIA Annual Symposium Proceedings*, vol. 2018, page 1524. American Medical Informatics Association.

Sikos, Jennifer and Sebastian Padó. 2019. Frame identification as categorization: Exemplars vs prototypes in embeddingland. In *Proceedings of the 13th International Conference on Computational Semantics - Long Papers*, pages 295–306. Gothenburg, Sweden: Association for Computational Linguistics.

Subirats, Carlos and Miriam R.L. Petruck. 2010. Surprise: Spanish FrameNet! *Estudios de Lingüística del Español* 31.

Tan, Sang-Sang and Jin-Cheon Na. 2019. Positional attention-based frame identification with BERT: A deep learning approach to target disambiguation and semantic frame selection. *arXiv preprint arXiv:1910.14549* .

Taniguchi, Ryosuke, Reina Hoshino, and Yoshinobu Kano. 2018. Legal question answering system using FrameNet. In *JSAI International Symposium on Artificial Intelligence*, pages 193–206. Springer.

Torrent, Tiago Timponi, Michael Ellsworth, CF Baker, and EE Matos. 2018. The multilingual FrameNet shared annotation task: a preliminary report. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, pages 62–68.

Torrent, Tiago Timponi, Maria Margarida Salomão, Fernanda Campos, Regina Braga, Ely Matos, Maucha Gamonal, Julia Gonçalves, Bruno Souza, Daniela Gomes, and Simone Peron. 2014. Copa 2014 FrameNet brasil: a frame-based trilingual electronic dictionary for the football world cup. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: System Demonstrations*, pages 10–14.

Trott, Sean, Tiago Timponi Torrent, Nancy Chang, and Nathan Schneider. 2020. (Re)construing meaning in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5170– 5184. Online: Association for Computational Linguistics.

Upadhyay, S., Faruqui, M., Dyer, C., & Roth, D. (2016). Cross-lingual models of word embeddings: An empirical comparison. In *54th Annual Meeting of the Association for Computational Linguistics, ACL 2016* (pp. 1661-1670). Association for Computational Linguistics (ACL).

Vossen, P.T.J.M., A.S. Fokkens, E. Maks, and C.M. van Son. 2018. Towards an open dutch framenet lexicon and corpus. In *Proceedings*, pages 75–80. ISBN 9791095546047. International FrameNet Workshop at LREC, May 12, 2018, LREC; Conference date: 12-05-2018 Through 12-05-2018.

Wu, Shijie and Mark Dredze. 2019. Beto, bentz, becas: The surprising cross-lingual effectiveness of BERT. 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 833–844. Hong Kong, China: Association for Computational Linguistics.