Learning to Align across Languages: Toward Multilingual FrameNet

Luca Gilardi, Collin F. Baker
International Computer Science Institute
1947 Center St., Berkeley, CA, 94704
{collinb,lucag}@icsi.berkelely.edu

Abstract

The FrameNet (FN) project, developed at ICSI since 1997, was the first lexical resource based on the theory of Frame Semantics, and documents contemporary English. It has inspired related projects in roughly a dozen other languages, which, while based on frame semantics, have evolved somewhat independently. Multilingual FrameNet (MLFN) is an attempt to find alignments between them all. The degree to which these projects have adhered to Berkeley FrameNet frames and the data release on which they are based varies, complicating the alignment problem. To minimize the resources needed to produce the alignments, we will rely on machine learning whenever that's possible and appropriate. We briefly describe the various FrameNets and their history, and our ongoing work employing tools from the fields of machine translation and document classification to introduce a new relation of similarity between frames, combining structural and distributional similarity, and how this will contribute to the coordination of the FrameNet projects, while allowing them to continue to evolve independently.

Keywords: frame semantics, cross-lingual resources, lexical resources, semantic roles

1. The FrameNet Project at ICSI

Developing tools and resources to move beyond the word or syntax level to the level of semantic analysis has long been a goal in natural language processing (NLP). In 1997, the FrameNet (FN) Project (Fillmore and Baker, 2010; Fontenelle, 2003) was started at the International Computer Science Institute (ICSI) `http://www.icsi.berkeley.edu`, initially funded by a three-year NSF grant, with the late Prof. Charles J. Fillmore as PI with the goal of establishing a general-purpose resource for frame semantic descriptions of English language text. FrameNet's lexicon is organized not around words, but **semantic frames** (Fillmore, 1976), which are characterizations of events, static relations, states, and entities. Each frame provides the conceptual basis for understanding a set of word senses, called **lexical units (LUs)**, that **evoke** the frame in the mind of the hearer; LUs can be any part of speech, although most are nouns, verbs, or adjectives. FrameNet now contains roughly 1,200 frames and 13,600 LUs.

FrameNet provides very detailed information about the syntactic-semantic patterns that are possible for each LU, derived from annotations on naturally occurring sentences. Annotators not only mark the frame-evoking LUs, but also label the phrases that instantiate the set of roles involved in the frame. These are known as **frame elements (FEs)**. An example of a simple frame is **Placing**, which represents the notion of someone or something placing something in a location. The core frame elements of **Placing** are the AGENT who does the placing (or the CAUSE of the placing), the THEME that is placed, and the GOAL. This is exemplified in annotated sentences containing LUs like *place.v, put.v, lay.v, implant.v*, and *billet.v* and also those like *bag.v*, *bottle.v*, and *box.v*, which already **incorporate** the GOAL, so that it need not be separately expressed. An example of a more complex frame is **Revenge**, which has FEs AVENGER, INJURED PARTY, INJURY, OFFENDER, and PUNISHMENT, as in

(1) [PUNISHMENT This book] is [AVENGER his] *REVENGE* [OFFENDER on his parents].

FrameNet semantic frames have been linked to form a densely connected lattice via eight different types of **frame relations**, including inheritance (subtype) relations and subparts of complex events.

FrameNet in NLP. FrameNet's main publications have been cited over 2,500 times according to Google Scholar, and the database, in XML format, has been downloaded thousands of times by researchers and developers around the world. Additionally, the well-known NLP library NLTK (Loper and Bird, 2002) provides API access to FrameNet.

Since FrameNet provides a uniquely detailed account of the syntactico-semantic patterns of use of a substantial number of common English words, there has been much interest in finding methods to annotate text automatically, using machine learning, training on the FrameNet data. The first system to use FrameNet for this purpose was developed by Daniel Gildea and Daniel Jurafsky (Gildea and Jurafsky, 2000). Automatic semantic role labeling has since become one of the standard tasks in NLP, and many freely available ASRL systems for FrameNet, have been developed. Recent systems include the SEMAFOR system developed at CMU by Dipanjan Das and colleagues (Das et al., 2010; Das et al., 2013). The latest semantic role labeling systems are able to improve accuracy by exploiting both FrameNet and PropBank jointly and also making use of the information from the frame hierarchy to produce FrameNet annotations ((FitzGerald et al., 2015; Kshirsagar et al., 2015; Roth and Lapata, 2015; Swayamdipta et al., 2017)). ASRL tools trained on FrameNet then enable a host of downstream NLP applications.

ASRL has also often been trained on PropBank,(Palmer et al., 2005) a resource inspired by FrameNet but specifically designed as an ASRL training corpus, without Fillmore's semantic frames. The term somewhat broader term *seman-*

tic parsing refers to the process of creating a semantic representation of a sentence or text; beside FrameNet-based ASRL, it has also been applied to systems aimed at creating formal logical representations.

2. FrameNet-related Projects for Other Languages

Since the beginning of Frame Semantics, the question arose as to whether semantic frames represent "universals" of human language or are language specific. While there are certainly many culturally specific phenomena and language-specific preferences in patterns of expression, the conclusion from the ICSI FrameNet experience has been that many frames can be regarded as applying across different languages, especially those relating to basic human experiences, like eating, drinking, sleeping, and waking. Even some cultural practices are similar across languages, such as commercial transactions: in every culture, commercial transactions involve the roles buyer, seller, money, and goods (or services).

Once the Berkeley FrameNet (hereafter BFN) project began releasing its data, researchers in many countries expressed interest in creating comparable resources for other languages. Despite the major effort required, a number of teams have persisted and been funded for substantial projects to create lexical databases for a wide variety of languages. Every FrameNet in another language constitutes an experiment in cross-linguistic Frame Semantics. The methods used in building these FrameNet have differed and each has created frames based on their own linguistic data, but all at least have an eye to how those frames compare with those created for English at ICSI (Boas, 2009). In the remainder of this section, we introduce the major FrameNets for languages other than English, and summarize some statistics for them in Table 1

Chinese FrameNet. The Chinese FrameNet Project ((You and Liu, 2005) `http://sccfn.sxu.edu.cn/`), based at Shanxi University in Taiyuan, was launched by Prof. Liu Kaiying in 2004, and is headed by Prof. Li Ru. It is based on the theory of Frame Semantics, making reference to the English FrameNet work in Berkeley, and supported by evidence from a large Chinese corpus. Currently, the Chinese FrameNet database contains 1,320 frames, 1,148 of the frames contain lexical units and 172 are non-lexical. There are 11,097 lexical units and nearly 70,000 sentences annotated with both syntactic and frame-semantic information. 3,616 of the LUs have annotated sentences; another 50,528 annotated sentences are being proofread and will be included in the database managing system. The lexicon covers both the common core of the language and the more specialized domains of law, tourism, and on-line book sales, as well as 200 discourses.

In addition to building the lexical database, the CFN team are studying the theory of frame semantics as it relates to the Chinese language, annotation of null instantiation, and extraction of Frame Semantic core dependency graphs for Chinese. They have developed frame semantic role labeling systems for both individual sentences and discourses (Li et al., 2010), and are researching techniques for building applications based on these. They have published more than 30 papers on Frame Semantics and building Chinese lexical resources.

Danish FrameNet Danish FrameNet (Nimb (2018) in this workshop, `https://github.com/dsldk/dansk-frame-net`) has been constructed by combining a Danish thesaurus and a Danish dictionary. The thesaurus has 1487 semantic groups which contain 42,000 words and expressions related to events (including intentional acts); these formed the starting point for the project. These were then connected to a dictionary which provided valence patterns for the words; on the basis of the valence patterns, the Danish words were translated into English and manually assigned to Berkeley FN frames, requiring 671 different frames. The researchers also studied which groupings in the thesaurus represent semantic domains not yet covered in Berkeley FrameNet. This project has apparently not done any annotation yet.

Dutch FrameNet The Dutch FrameNet project ((Vossen et al., 2018) in this workshop, `https://github.com/cltl/Open-Dutch-Framenet`) started from a Dutch corpus with PropBank annotations and annotated 5,250 tokens of 1,335 verb lemmas that were already selected during the annotation of the PropBank values. Only the main verb of the sentence and its arguments were annotated with a frame an its frame elements. All other verbs (such as auxiliaries and modals) and all other parts-of-speech were left unannotated for the present, along with nouns and adjectives. These represented 4,755 LUs in 671 frames, all chosen from Berkeley FN. All were annotated by two researchers. They adopted an unusual policy with respect to disagreement between annotators– they kept both annotations, rather than asking an expert to adjudicate between them. Because they were working from corpus data rather than a list of lexical items, all of the lemmas in the lexicon have at least some annotated examples.

Finnish FrameNet Finnish FrameNet (Lindén et al. (2017), `http://urn.fi/urn:nbn:fi:lb-2016121201`), was created on a frame-by-frame basis, using the BFN frames. First, some 80,000 sentences from Berkeley FrameNet were chosen and the parts of the sentence which had been annotated in English were professionally translated to Finnish, creating an "English-Finnish TransFrame Corpus". Then Finnish newspaper articles were searched for sentences with similar syntax and semantics, and these were manually annotated. The researchers found that it was necessary to change the annotation practices from those of BFN, and annotate the morphemes within words in Finnish, as might be expected given the agglutinative nature of Finnish. However, the principal result of the experiment was the finding that in most cases, the English frames generalized well to Finnish, even though it is a completely unrelated language with very different morphology and syntax.

FrameNet Brasil FrameNet Brasil ((Torrent et al., Forthcoming; Torrent et al., 2014) `http://www.framenetbr.ufjf.br`) has been one of the most active and productive FrameNets in recent years, producing both theoretic insights and practical, real-world applications of Frame Semantics. It is also the only project that has created a multilingual FrameNet internally.

FrameNet Brasil started in 2007 and the first data release was in 2010. The project is headquartered in the Computational Lexicography Lab at the Federal University of Juiz de Fora, Minas Gerais. There are two main lines of development, one of which is focused on creating a Brazilian Portuguese parallel to ICSI FrameNet, together with an integrated "Constructicon". The other line is building frame-based domain-specific multilingual applications for non-specialist users, which began with the creation of the FrameNet Brasil World Cup Dictionary (www.dicionariodacopa.com.br), a dictionary for the 2015 Soccer World Cup containing 128 frames and over 1,000 lexical units, in English, Portuguese, and Spanish. The main development is now on the successor application, the Multilingual Knowledge Base (m.knob), a trilingual travel assistant app that offers personalized information to tourists about the specific domains of Tourism and Sports. The alpha version of the app was released during the Rio 2016 Summer Olympics and has been redesigned to include other functions in its beta version. M.knob has two main functions, (i) a chatbot providing recommendations on tourist attractions and activities; and (ii) a semantically enhanced sentence translator algorithm based on frames and qualia relations (Pustejovsky, 1995). These functions have required creation of many new frames in the sports and tourism domains; m.knob currently features 58 frames for tourism and sports, only 16 of which already existed in the Berkeley FrameNet Data Release 1.7. For the Sports Domain, Costa and Torrent (2017) created 29 new frames and used 4 frames from Berkeley FrameNet 1.7. Currently, the m.knob lexicon comprises a total of 5,152 LUs: 1,671 for Brazilian Portuguese, 2,551 for English, 930 for Spanish (da Costa et al. (2018) in this workshop). Texts were extracted from travel guides and blogs, governmental portals on tourism and on the Olympics, as well as from sports manuals and websites of associations of each Olympic sport.

The need to model these domains in multiple languages and to model constructions fully in the same database as semantic frames has led to changes in database structure which permit creation of new relations and new kinds of relations between fields in the database which are not connected in Berkeley FrameNet. Space limits prohibit discussing these changes fully here, but we can note that the new FN Brasil database allows one to freely create relations between any two objects in the database.

French FrameNet. French FrameNet, (Candito et al. (2014) https://sites.google.com/site/anrasfalda/ which operated from October 2012 to June 2016) was headed by Prof. Marie Candito, with about 15 researchers at three sites, U Paris Diderot, Toulouse, and Aix-Marseille, as well as industrial partners, and was set up within the ASFALDA project, funded by ANR and the Empirical Foundations of Linguistics Labex. French FrameNet focuses on four notional domains (verbal communication, commercial transactions, cognitive stance, and causality). The objective of the project was to exhaustively cover these four domains, in terms of relevant frames, lexical units and annotation. They performed manual annotation domain by domain, on two pre-existing syntactic treebanks, the French Treebank (Abeillé and Barrier, 2004) and the Sequoia Treebank (Candito and Seddah, 2012). Release 1.3 of French FrameNet contains 106 frames, 1,936 lexical units and 16,167 annotation sets. Among their frames, roughly 60% are the same as those of English FrameNet Release 1.5, 13 % are modified English frames, 11% were created by splitting English frames, 7% were created by merging English frames, and 9% are new frames. The annotation style also differs somewhat from English FrameNet, in that most non-core frame elements of verbs are not annotated; instead, prepositions and conjunctions are annotated as frame-evoking elements, to represent similar semantic relations.

German FrameNet research The SALSA project ((Burchardt et al., 2006; Burchardt et al., 2009a), http://www.coli.uni-saarland.de/projects/salsa) from 2002 to 2010 in Saarbrücken, Germany under the direction of PI Manfred Pinkal, explored methods for large-scale manual frame-semantic annotation of entire news stories from the German TIGER Treebank (Brants et al., 2002), and multilingual approaches to inducing and verifying frame semantic annotations. The annotators used the English FN frames where possible, but when they ran into words for which there was no corresponding LU in ICSI FrameNet, they created "proto-frames", i.e. provisional frames for a single lexeme, without grouping them into larger frames. The second release of the SALSA annotated corpus is freely available.

The Saarbrücken team also did research on using frame semantic annotation to help with the textual entailment task (Burchardt et al., 2009a) and released a freely available training corpus for this purpose (Burchardt and Pennacchiotti, 2008; Burchardt et al., 2009b).

Recently, there has been renewed interest in creating a larger German FrameNet, possibly based on the work of SALSA. A group of German researchers have begun a collaborative exchange program with FrameNet Brasil, and Prof. Oliver Czulo of University of Leipzig has set up a project do full-text annotation of the German version of the TED talk "Do Schools Kill Creativity?"; this is part of a larger annotation project, done in parallel with other FrameNets, to be discussed later in this workshop (Torrent et al., 2018). They are using the WebAnno tool. In addition, a conference on "Issues in Multilingual Frame Semantics: Comparability of frames" will be held in October at University of Leipzig, which will deal with comparability of German frames, *inter alia.* They are also working on a "constructicon", first for German, but later for English (www.german-constructicon.de[www.german-constructicon.de). Also, Prof. Hans Boas, at University of Texas at Austin is leading work on manual lexical annotation of the online first-year German textbook "Deutsch im Blick", building up a frame semantic dictionary of German as a second language (Boas et al. (2016), http://coerll.utexas.edu/frames/home).

Hebrew FrameNet Hebrew FrameNet (Hayoun and Elhadad, 2016) is being built at Ben-Gurion University of the

Negev by Prof. Michael ELhadad and (currently) grad student Ben Eyal. They have collected a database of roughly 23 million English-Hebrew sentence pairs from the Open Subtitles database and word-aligned and parsed both languages. They used the aligned 115 million aligned words as a bilingual dictionary to translate English LUs to produce 5258 Hebrew LUs. They then run the SEMAFOR automatic semantic role labeling system trained on FrameNet Release 1.5 over the English and create FE labeling on the Hebrew by projection to the equivalent constituents. In this way they have produced 11k automatically annotated sentences in 678 frames, and are in the process of manually verifying them. They are working on better automatic ways of finding example sentences for the LUs, search diversification (Borin et al., 2012), and of finding exemplar sentences for frames.

Hindi/Urdu FrameNet (Virk and Prasad, 2018)
Shafqat Mumtaz Virk and K. V. S. Prasad have just begun a new project to produce both Hindi and Urdu FrameNets. Since these are either closely related languages or somewhat distant dialects of the same language (depending on one's point of view), it will no doubt be advantageous for this research to be carried out jointly, and the similarities and differences documented will be instructive both theoretically and practically for other pairs of related languages. The main reference for the project is the paper and accompanying poster at this workshop (Virk and Prasad, 2018); they are planning to set up a website for the project soon.

At the moment, they are concentrating on full-text annotation of the TED talk; they actually had to produce the Hindi version themselves, since it did not exist when they began work. They consulted the English and Portuguese annotation of the talk as a reference. In some cases, the frames used there were acceptable for Hindi or Urdu, but in many cases, they were obviously not (as when the words of the translation evoke different images). In these latter cases, they annotated as best they could from scratch, noting the required changes in frame-structure and/or frame-elements for future Hindi/Urdu FrameNets. This strategy allowed them to get started quickly, but they plan to revisit the entire text later with no reference to previous annotations in other languages, to avoid distorting the frames towards previously created FrameNets.

Italian Frame Semantic Research. Researchers at Fondazione Bruno Kessler (FBK) and at the University of Trento have done a great deal of research on FrameNet. They began working on an Italian FrameNet in 2007, using a combination of manual annotation and automatic expansion and projection (Tonelli and Giuliano, 2009; Tonelli and Pianta, 2008) and concluded that "Italian frames only needed minimal adjustments to be imported from English..." They have used several techniques to expand the FrameNet lexicon (Tonelli and Pianta, 2009; Bryl et al., 2012). In the last of these, they also released a version of the FrameNet hierarchy in RDF notation as linked open data on the cloud.

Another group headed by Prof. Alessandro Lenci of University of Pisa (ILC–CNR) has used the English FN frames to annotate Italian verbs and tested a variety of semi-automatic techniques (Lenci et al., 2010).

Japanese FrameNet. The Japanese FrameNet Project was launched in 2002 ((Ohara et al., 2004), Ohara (2012), `http://jfn.st.hc.keio.ac.jp`); since 2005, it has been developed at Keio University, in cooperation with ICSI. Their annotated frames are imported from BFN and their database has the same structure as the ICSI one. Because they imported many BFN frames and translated many BFN LUs initially, they have a number of frames and LUs without annotation. Currently, they are annotating texts from the Balanced Corpus of Contemporary Written Japanese (BCCWJ) core data in collaboration with the National Institute for Japanese Language and Linguistics, and have also been building a "**constructicon**", a repertoire of grammatical constructions.

The Japanese FrameNet team has recently begun participating in a joint project at the RIKEN Center for Advanced Intelligence Projects; other members include Prof. Kawahara (Kyoto University), Kentaro Inui (Tohoku University), and Satoru Sekine (New York University); they are working on scaling up Japanese FrameNet using crowdsourcing. The early crowdsourcing results are providing indications of which specific LUs/annotations should be corrected or added to.

Korean FrameNet Korean FrameNet (`http://framenet.kaist.ac.kr/`) has been created in part by using expert translations of annotated sentences from the Berkeley and Japanese FrameNets into Korean, projecting the FE annotation to corresponding constituents in Korean (Hahm et al., 2014). They have also translated LU names into Korean, giving them more than 8000 LUs, but many are not annotated. They have calculated the coverage of basic Korean vocabulary and studied the valence patterns, comparing English to Korean valences for similar verbs. They are currently linking Korean WordNet to English WordNet and then (via WordNet to FrameNet mappings) to FrameNet frames. They are using the resulting database for Frame Semantic parsing of Korean; their goal is to annotate the 300k articles of the Korean Wikipedia (K.S. Choi, p.c.).

Latvian FrameNet Latvian FrameNet (`https://github.com/LUMII-AILab/FullStack`) is using a corpus-driven approach; the input text is parsed using Universal Dependencies (Gruzitis et al. (2018a) in this workshop), and then annotated with FrameNet FEs using WebAnno 3 (https://webanno.github.io/webanno/), with some customization. Because the dependency structure is available, the annotator marks only the head of the phrase, depending on the parse for the ends of the span; this is similar to the approach used in SALSA (and many other annotation projects). Their internal data format is flat tables, similar to CoNLL.

The annotation is similar to BFN "lexicographic" annotation, annotating many sentences for only one LU, although the same sentence can be reused for another LU; they are not yet doing full-text annotation. For the moment, they are keeping to the BFN Release 1.7 fame inventory; when no appropriate frame can be found, they use a more general one. An earlier project built a Latvian FrameNet spe-

Project	Total Frames	Total LUs	Total Anno. Sets
FrameNet (ICSI)	1,224	13,639	202,229
Chinese FN	320	3,200	22,000
Danish FN	671	33,930	0
Dutch FN	671	4755	5250
Finnish FN	938	6,639	40,721
FN Brasil (PT)	472	2896 ‘x	11,779
FN Brasil m.knob (PT)	91	1671	7912
FN Brasil m.knob (EN)	91	350	3374
FN Brasil m.knob (ES)	91	360	2398
French FN (Asfalda)	96	727	10,632
German FN (SALSA)	1,023 (768)	650	37,697
Hebrew FN	157	5258	11,205
Hindi FN	84	84	?
Italian FN	38	211	–
Japanese FN	979	5029	7899
Korean	722	8220	5507 sents.
Latvian	319	1350	10334
Spanish FN	325	1,350	10,334
Swedish FN++	1,215	39,558	9,223
Urdu FN	42	42	?

Table 1: Summary of FrameNet Projects by Language

cific to the news domains using a controlled natural language approach for NLU Barzdins (2014) and NLG (Gruzitis and Dannélls, 2017). The current project is intended to be part of a larger multi-layer representation including an Abstract Meaning representation (AMR) layer (Gruzitis et al., 2018b).

Spanish FrameNet. Spanish FrameNet (SFN) ((Subirats, 2009), `http://spanishfn.org/`) is being developed at the Autonomous University of Barcelona under the direction of Carlos Subirats, with colleagues at ICSI and throughout Spain. When they began work in 2002, they found that there was no suitable balanced corpus of Spanish which reflected the importance of New World Spanish, so they put together their own corpus. They also created their own POS tagging system. Because their practices have remained close to the Berkeley model, they were able to use a minor modification of the ICSI tools for corpus search and visualization of the frame hierarchy. Their annotated lexicographic examples have also been used to train automatically semantic role labelers for Spanish text.

Swedish FrameNet++. The Swedish FrameNet project (SweFN++, (Borin et al., 2010), `https://spraakbanken.gu.se/eng/swefn`) was developed in the Språkbanken NLP research group at U. Gothenburg. The main purpose of Swedish FN was to make a framenet available for Swedish NLP; therefore, they have reused the BFN frames and simply populated them with Swedish LUs, resulting in a very large lexicon, but have not tried to annotate a large number of corpus examples. They have, however added new frames for Swedish LUs which did not fit into any existing BFN frame.

The other objective of the project was to integrate a large and varied collection of computational lexical resources, including SALDO,(Borin et al., 2013), a large morphological and lexical-semantic lexicon for modern Swedish, using a uniform identifier format for word senses (i.e., FN LUs), inflectional units, sense relations, etc. and supplement them with FN frames (hence the “++” in the name). This part enables them to draw on framenet information elsewhere, for example in their historical lexicons.

The SweFN team have collaborated extensively in the development of FrameNets in new languages and specialized domains. They are currently in collaboration with FrameNet Brasil, and helping with the creation of new FrameNet projects for Hindi and Urdu. The latest downloadable version of the SweFN data is at `https://svn.spraakdata.gu.se/sb-arkiv/pub/lmf/swefn/swefn.xml`.

Other recent FrameNets There have been recent efforts on many other languages, and keeping up with them has become difficult. Here are some which we know about: Slovenian (Lönneker-Rodman et al., 2008), Bulgarian (Koeva, 2010), and Polish ((Zawisławska et al., 2008), `http://www.ramki.uw.edu.pl/en/index.html`).

Table 1 shows summary statistics for most of the FrameNets discussed above. The numbers shown here are gleaned from a variety of websites, papers, and personal communications and represent our best estimates, but may not be current in all cases. We apologize in advance if we have incorrect figures for any of the projects. The counts for frames represent Berkeley FrameNet frames in most cases, but as discussed above, certain projects, such as the Brazilian m.knob project, have created domain specific frames which have not been incorporated into BFN; and different projects have used more automatic or more manual methods of creating LUs and annotating to sentences, so the numbers are often not directly comparable.

3. Towards an Aligned Multilingual FrameNet

3.1. Overview

Given that so much research has been conducted in building separate lexical databases for many languages using a set of semantic frames that are largely the same across languages, it is natural to ask whether these lexical databases could be aligned to form a multilingual FrameNet lexical database connecting all of the languages mentioned above, as well as others in the future, and whether this can be done while also accounting for language-specific differences and domain-specific extensions to FrameNet. The results of work done during the planning phase suggest that both of these task are possible. We also feel that it is urgent to carry out this harmonization process as soon as possible, to take better advantage of the experience of each language project, to avoid duplication of effort, and to unify the representational format as much as possible.

Despite differences among the various FrameNet projects discussed above, all agree on the concept of semantic frames as the organizing principle of their lexicons and in general all have found the set of frames defined in the Berkeley project sufficiently general to be widely applicable to their language. On the other hand, the differences in the degree to which the projects have adhered to Berkeley FrameNet (BFN) complicate the alignment problem. The Spanish, Japanese, and Brazilian FNs have followed BFN rather closely, using BFN frames as templates, whereas the SALSA Project, Swedish FrameNet++ and Chinese FN have allowed a greater degree of divergence from BFN, either adding many new frames and/or modifying the BFN-derived ones. (At this time, the MLFN effort is not trying to align the French, Italian or Hebrew efforts, for various reasons, which include availability, coverage, and other aspects.)

More specifically, divergence of approaches means that we also need different approaches to the alignment task. For the first group, we can largely rely on BFN's frame elements and IDs, and use an algorithm roughly like the following:

- for each pair of projects (BFN, XFN):
  - Compare each individual Lexical Unit in each BFN Frame with each lexical unit in the corresponding XFN frame
  - Compare the frame definitions, FEs, Semantic Types, and Relations

For each comparison, we need a metric to assess the similarity. Such a metric has to take into account that if, for example, two frames with the same name have different sets of core FEs, strictly speaking, they should not be considered the same frame. One possible metric might be built on a variant of the Jaccard Index, which is used to identify similarity between sets, attributes, or vectors. For the second group, the alignment process is not so straightforward; for some frames, we either assume that they have no overlap with any frame in BFN, or we try to find some relatively closely corresponding frame in BFN, by using the same similarity metric as for the first group, but applied to every possible cross-lingual pair of frames.

An additional complication arises because even the projects that strictly adhere to BFN have branched off at different times, and were based on different versions of BFN: for example, Spanish FN was based on BFN Release 1.5, others on Release 1.2. Thus, we need to:

- Find a mapping back from the current BFN to the BFN version used by the project at hand (let's call it xFN)
- Find a mapping from the earlier ICSI FrameNet version to x_{FN}
- And then compose the two mappings

A further twist is that in some cases, projects developing in parallel (such as SALSA and BFN) have influenced each other, often adding very similar, but not identical frames. All of this suggests that it would be helpful if MLFN had a way to track such interactions over time. Such a feature should be included in future versions on the FN database management software.

We have built support software which allows each data from each project to be directly imported in its native format (typically, XML files, but also SQL data), but the problem of maintaining a growing MLFN database remains. In order to minimize the collaborative effort requiree in the construction of a lexical resource like FrameNet, it would be desirable to retrofit the MLFN management software with a versioned database, i.e. one that makes it possible, for any language, to track and control the revisions of frames, FEs, LUs, and the relations among them, i.e. incorporating features analogous to those of the version control systems used to manage revisions of software and documentation.

3.2. Aligning FrameNets

In planning Multilingual FrameNet, we assume that more projects in new languages will be added in the future, and that it is therefore advisable to minimize the amount of human effort needed to integrate new projects and maintain the overall structure of the MLFN project.

The current alignment effort focuses less on infrastructure and more on the direct applicability of the deliverables, and relies on statistical methods where possible. We can evaluate the progress of this effort in two different ways: either in the abstract, locating and quantifying differences in frames and FEs in different projects, or more concretely, measuring the effect of those differences on a common computational task that uses FN as a component.

The core of the MLFN alignment algorithm proceeds in a pairwise fashion by matching and afterwards aligning BFN with each of the other FrameNets. It has been devised in part by operationalizing some of ICSI's internal methods to avoid the creation of multiple frames, and by introducing a *weighted voting model*. This assumes that we have available a relatively reliable and accurate machine translation (MT) method between the two languages. The basic idea is to use it to generate LU-to-LU translations links to select possible frames for alignment. Thus, broadly speaking, we can say that a frame X is *aligned with* a frame Y to the extent that there are pairs of LUs associated with each frame that are good translations of each other. Since we want to take into consideration the errors made by the MT system, we will configure such system to output a list of possible translations for each input LU, together with their probabilities, and from that list generate *votes*, with associated weights computed from the probabilites.

Let us describe the process in a little more detail: For each matching pair $\langle E_{\text{FN}}, X_{\text{FN}} \rangle$ of English and non-English FN FrameNets, and for each Lexical Unit and frame E in E_{FN}, we find zero or more corresponding LUs and their frames X in X_{FN} by (automatically) translating the LU in the source language e to the target language x. We create a correspondence between the frames E and X, with each pair of LUs contributing a weighted *vote* to each such alignment. We then normalize (by the number of pairs of LUs that translate to each other) to obtain a weight w_{ij}, where i is an index over the frames in E_{FN} and j over those in X_{FN}, and add a new weighted relation between E and X. We call this new relation the **alignment** between frame E and X. By repeating this process for all pairs of languages, we can generate alignments between each pair of FrameNet projects. The new relation, ALIGNED_WITH, is a weighted arc, which is unusual for FrameNet, but necessary because not all frames in different projects overlap perfectly, and also because MLFN cannot assume that such overlap is even possible in all cases (e.g., some frames are culture-specific frames, some others encode semantics that would be better captured by constructions, etc. (ae shown in (Ohara, 2008)).

As already noted, while the proposed alignment method tries to mitigate the effect of possible mistranslations between the Lexical Units in different languages, it still depends crucially on some form of automatic translation. We are considering several possibilities: a simple translation based on a dictionary with word senses can be used as a baseline, or, for instance, one based on Open Multilingual WordNet (Bond and Paik, 2012) or the UWN/MENTA project (de Melo and Weikum, 2009).

Ideally though, we would like to employ methods that take into account the syntactic and semantic environment in which words are used. One option that is increasingly popular is to use distributional representations such as Word2Vec (Mikolov et al., 2013) or GloVe (Pennington et al., 2014). A more recent study also shows how to learn alignments from monolingual word vectors in 98 languages (Smith et al., 2017). Although these methods do not try to explicitly encode syntactic relations, some others do: for example, (Pado and Lapata, 2007) show how to generate vector representation starting from dependency parses.

These methods work for language pairs, which entails that each pair of FN projects would need specific training data and computational resources. Moreover, the methods described in (Pado and Lapata, 2007) require the availability of syntactic parsers for all the languages involved; this might be a problem n some case, since not all languages have NLP resources like those for English.

But even without considering syntactic parsing, adding a new language to MLFN thus would require separate training for each of the languages already in place. Fortunately, the MT community has for some time been developing vector representations specifically geared towards multilingual environments; these vectors in joint (cross-lingistic) spaces make it possible, for instance, to translate from French to German having only trained the system with parallel corpora pairs English-German and English-French. For a small survey of these methods, see e.g. (de Melo, 2017).

Hermann and Blunsom (2014; Søgaard et al. (2015) describe methods that, starting from multilingual parallel corpora, not only generate semantic vectors that jointly represent multiple languages in the same semantic space, but also encode additional information about the larger context in which the LUs are used—the document context in their case, since they evaluate their vector representations in a document classification task. We plan to implement a similar approach, along the lines of (Hermann and Blunsom, 2014), in which the larger context is instead the set of frames in which word forms appear.

We are currently studying methods for separately learning the joint-space representations of words from parallel corpora, and from (ML)FrameNet annotations to investigate the relations between vector representations and frames. Thus we hope that our research will yield a compositional method to relate joint-space representations of words to frames. The rationale is that we would like to use the richer resource to learn frame assignments, and then transfer these learned relations to FrameNet projects that have fewer annotations, or no annotations at all; in this way we might be able to help to jump-start new FrameNet projects for low-resource languages.

We plan to evaluate our system in a Multilingual Frame Identification task. In Semantic Role Labeling (SRL) systems (e.g. (Gildea and Jurafsky, 2002; Das et al., 2013; Roth and Lapata, 2015; Swayamdipta et al., 2017)), the process is usually divided into two subtasks: (i) Frame Identification (FI), and (ii) Argument Identification The latter assumes that a suitable frame for the target has been found and proceeds to attach FE names to the relevant arguments. Therefore argument identification relies crucially on the FI phase. By providing multilingual FI capabilities we would also be enabling the implementation of SRL systems based on MLFN.

4. Conclusions

To summarize, our alignment scheme offers a unified view of the different FrameNet projects, which includes weighted relations between the frames in all the projects, a frame similarity metric both across projects and within the same project, a Frame Identification tool to suggest possible frame assignments for LUs that are present in some projects and absent in others, and utilities for importing projects in their native format. We plan to make the Multilingual FrameNet database, algorithms, training and evaluation data available on-line in the next few months.

5. Acknowledgments

This material is based in part upon work supported by the U.S. National Science Foundation under Grant No. (1629989). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the

authors and do not necessarily reflect the views of the National Science Foundation.

6. Bibliographical References

Abeillé, A. and Barrier, N. (2004). Enriching a french treebank. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*. European Language Resources Association (ELRA).

Barzdins, G. (2014). Framenet CNL: A knowledge representation and information extraction language. In *International Workshop on Controlled Natural Language*, pages 90–101. Springer.

Boas, H. C., Dux, R., and Ziem, A. (2016). Frames and constructions in an online learner's dictionary of german. In S. De Knop et al., editors, *Applied Construction Gammar*, pages 303–326. de Gruyter.

Hans C. Boas, editor. (2009). *Multilingual FrameNets in Computational Lexicography: Methods and Applications*. Mouton de Gruyter.

Bond, F. and Paik, K. (2012). A survey of wordnets and their licenses. In *Proceedings of the 6th Global WordNet Conference (GWC 2012)*, Matsue. 64–71.

Borin, L., Danélls, D., Forsberg, M., Kokkinakis, D., and Gronostaj, M. T. (2010). The Past Meets the Present in Swedish FrameNet++. In *Proceedings of EURALEX 14*, pages 269–281. EURALEX.

Borin, L., Forsberg, M., Friberg Heppin, K., Johansson, R., and Kjellandsson, A. (2012). Search result diversification methods to assist lexicographers. In *Proceedings of the Sixth Linguistic Annotation Workshop*, pages 113–117. Association for Computational Linguistics.

Borin, L., Forsberg, M., and Lönngren, L. (2013). SALDO: a touch of yin to WordNet's yang. 47(4):1191–1211.

Brants, S., Dipper, S., Hansen, S., Lezius, W., and Smith, G. (2002). The TIGER Treebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories*, Sozopol.

Bryl, V., Tonelli, S., Giuliano, C., and Serafini., L. (2012). A Novel FrameNet-based Resource for the Semantic Web. In *Proceedings of ACM Symposium on Appliced Computing (SAC)*, Riva del Garda (Trento), Italy.

Burchardt, A. and Pennacchiotti, M. (2008). FATE: a FrameNet-Annotated Corpus for Textual Entailment. In *Proceedings of LREC 2008*.

Burchardt, A., Erk, K., Frank, A., Padó, S., and Pinkal, M. (2006). The SALSA Corpus: a German Corpus Resource for Lexical Semantics. In *Proceedings of 5th International Conference on Language Resources and Evaluation*, Genoa, Italy.

Burchardt, A., Erk, K., Frank, A., Kowalski, A., Padó, S., and Pinkal, M. (2009a). Using FrameNet for the semantic analysis of German: Annotation, representation, and automation. In Hans C. Boas, editor, *Multilingual FrameNets in Computational Lexicography*, pages 209–244. Mouton.

Burchardt, A., Pennachiotti, M., Thater, S., and Pinkal, M. (2009b). Assessing the impact of frame semantics on textual entailment. *Natural Language Engineering*, 15(Special Issue 04):527–550.

Candito, M. and Seddah, D. (2012). Le corpus Sequoia: annotation syntaxique et exploitation pour l'adaptation d'analyseur par pont lexical. In *TALN 2012-19e conférence sur le Traitement Automatique des Langues Naturelles*.

Candito, M., Amsili, P., Barque, L., Benamara, F., De Chalendar, G., Djemaa, M., Haas, P., Huyghe, R., Mathieu, Y. Y., Muller, P., et al. (2014). Developing a french framenet: Methodology and first results. In *LREC-The 9th edition of the Language Resources and Evaluation Conference*.

Costa, A. D. and Torrent, T. T. (2017). A modelagem computacional do domínio dos esportes na FrameNet Brasil (the computational modeling of the sports domain in FrameNet Brasil)[in portuguese]. In *Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology*, pages 201–208.

da Costa, A. D., Gamonal, M. A., Paiva, V. M. R. L., Natália Duarte Mar c. a., Peron-Corrêa, S. R., de Almeida, V. G., da Silva Matos, E. E., and Torrent, T. T. (2018). Framenet-based modeling of the domains of tourism and sports for the development of a personal travel assistant application. In *Proceedings of the International FameNet Workshop 2018: Multilingual FrameNets and Constructions*.

Das, D., Schneider, N., Chen, D., and Smith, N. A. (2010). Probabilistic Frame-Semantic Parsing. In *Proceedings of the North American Chapter of the Association for Computational Linguistics Human Language Technologies Conference*, Los Angeles, June.

Das, D., Chen, D., Martins, A. F. T., Schneider, N., and Smith, N. A. (2013). Frame-Semantic Parsing. *Computational Linguistics*, 40(1).

de Melo, G. and Weikum, G. (2009). Towards a universal wordnet by learning from combined evidence. In David Wai-Lok Cheung, et al., editors, *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM 2009)*, pages 513–522, New York, NY, USA. ACM.

de Melo, G. (2017). Multilingual vector representations of words, sentences, and documents. In *Proceedings of the IJCNLP 2017, Tutorial Abstracts*, pages 3–5. Asian Federation of Natural Language Processing.

Fillmore, C. J. and Baker, C. F. (2010). A Frames Approach to Semantic Analysis. In Bernd Heine et al., editors, *Oxford Handbook of Linguistic Analysis*, pages 313–341. OUP. This is Chapter 13 in 1st edition, 33 in 2nd edition.

Fillmore, C. J. (1976). Frame semantics and the nature of language. *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*, 280(1):20–32.

FitzGerald, N., Täckström, O., Ganchev, K., and Das, D. (2015). Semantic role labeling with neural network factors. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 960–970, Lisbon, Portugal, September. Association for

Computational Linguistics.

Thierry Fontenelle, editor. (2003). *International Journal of Lexicography–Special Issue on FrameNet*, volume 16. Oxford University Press.

Gildea, D. and Jurafsky, D. (2000). Automatic Labeling of Semantic Roles. In *ACL 2000: Proceedings of ACL 2000, Hong Kong*.

Gildea, D. and Jurafsky, D. (2002). Automatic labeling of semantic roles. *Comput. Linguist.*, 28(3):245–288, September.

Gruzitis, N. and Dannélls, D. (2017). A multilingual FrameNet-based grammar and lexicon for controlled natural language. *Language Resources and Evaluation*, 51(1):37–66.

Gruzitis, N., Nespore-Berzkalne, G., and Saulite, B. (2018a). Creation of Latvian FrameNet based on universal dependencies. In Tiago Timponi Torrent, et al., editors, *Proceedings of the International FameNet Workshop 2018: Multilingual FrameNets and Constructions*, Miazaki, Japan.

Gruzitis, N., Pretkalnina, L., Saulite, B., Rituma, L., Nespore-Berzkalne, G., Znotins, A., and Paikens, P. (2018b). Creation of a balanced state-of-the-art multilayer corpus for NLU. In *Proceedings of LREC 2018*.

Hahm, Y., Kim, Y., Won, Y., Woo, J., Seo, J., Kim, J., Park, S., Hwang, D., and Key-Sun-Choi. (2014). Toward matching the relation instantiation from DBpedia ontology to Wikipedia text: Fusing FrameNet to Korean. In *Proceedings of the 10th International Conference on Semantic Systems*, pages 13–19.

Hayoun, A. and Elhadad, M. (2016). The Hebrew FrameNet Project. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Paris, France, may. European Language Resources Association (ELRA).

Hermann, K. M. and Blunsom, P. (2014). Multilingual models for compositional distributed semantics. *CoRR*, abs/1404.4641.

Koeva, S. (2010). Lexicon and Grammar in Bulgarian FrameNet. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

Kshirsagar, M., Thomson, S., Schneider, N., Carbonell, J., Smith, N. A., and Dyer, C. (2015). Frame-Semantic Role Labeling with Heterogeneous Annotations. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 218–224, Beijing, China, July. Association for Computational Linguistics.

Lenci, A., Johnson, M., and Lapesa, G. (2010). Building an Italian FrameNet through Semi-automatic Corpus Analysis. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta, may. European Language Resources Association (ELRA).

Li, J., Wand, R., and Gao, Y. (2010). Sequential tagging of semantic roles on Chinese FrameNet. In *Proceedings of the Eighth Workshop on Asian Language Resouces*, pages 22–29. Coling 2010 Organizing Committee.

Lindén, K., Haltia, H., Luukkonen, J., Laine, A. O., Roivainen, H., and Väisänen, N. (2017). FinnFN 1.0: The Finnish frame semantic database. *Nordic Journal of Linguistics*, 40(3):287–311.

Lönneker-Rodman, B., Baker, C., and Hong, J. (2008). The New FrameNet Desktop: A Usage Scenario for Slovenian. In Jonathan Webster, et al., editors, *Proceedings of The First International Conference on Global Interoperability for Language Resources*, pages 147–154, Hong Kong. City University.

Loper, E. and Bird, S. (2002). NLTK: The natural language toolkit. In *In Proceedings of the ACL Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics. Philadelphia: Association for Computational Linguistics*.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *CoRR*, abs/1301.3781.

Nimb, S. (2018). The Danish Framenet lexicon: Method and lexical coverage. In *Proceedings of the International FameNet Workshop 2018: Multilingual FrameNets and Constructions*, pages 48–52.

Ohara, K., Fujii, S., Ishizaki, S., Ohori, T., Saito, H., and Suzuki, R. (2004). The Japanese FrameNet Project; An introduction. In Charles J. Fillmore, et al., editors, *Proceedings of the Workshop on Building Lexical Resources from Semantically Annotated Corpora*, pages 9–12, Lisbon. LREC 2004.

Ohara, K. (2008). Lexicon, grammar, and multilinguality in the Japanese FrameNet. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May.

Ohara, K. (2012). Semantic Annotations in Japanese FrameNet: Comparing Frames in Japanese and English. In Nicoletta Calzolari (Conference Chair), et al., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Pado, S. and Lapata, M. (2007). Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2):161–199.

Palmer, M., Gildea, D., and Kingsbury, P. (2005). The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71–106, March.

Pennington, J., Socher, R., and Manning, C. D. (2014). Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Roth, M. and Lapata, M. (2015). Context-aware Frame-Semantic Role Labeling. *Transactions of the Association for Computational Linguistics*, 3:449–460.

Smith, S. L., Turban, D. H. P., Hamblin, S., and Hammerla, N. Y. (2017). Offline bilingual word vectors, orthogonal transformations and the inverted softmax. *CoRR*, abs/1702.03859.

Søgaard, A., Agic, Z., Alonso, H. M., Plank, B., Bohnet, B., and Johannsen, A. (2015). Inverted indexing for cross-lingual NLP. In *ACL*.

Subirats, C. (2009). Spanish FrameNet: A frame-semantic analysis of the Spanish lexicon. In Hans Boas, editor, *Multilingual FrameNets in Computational Lexicography: Methods and Applications*, pages 135–162. Mouton de Gruyter, Berlin/New York.

Swayamdipta, S., Thomson, S., Dyer, C., and Smith, N. A. (2017). Frame-semantic parsing with softmax-margin segmental RNNs and a syntactic scaffold.

Tonelli, S. and Giuliano, C. (2009). Wikipedia as frame information repository. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1 - Volume 1*, EMNLP '09, pages 276–285, Stroudsburg, PA, USA. Association for Computational Linguistics.

Tonelli, S. and Pianta, E. (2008). Frame Information Transfer from English to Italian. In European Language Resources Association (ELRA), editor, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may.

Tonelli, S. and Pianta, E. (2009). A novel approach to mapping FrameNet lexical units to WordNet synsets. In *Proceedings of IWCS-8*, Tilburg, The Netherlands, January.

Torrent, T. T., Salomão, M. M. M., and Peron, S. R. (2014). Copa 2014 FrameNet Brasil: a frame-based trilingual electronic dictionary for the Football World Cup. In *Proceedings of 25th COLING: System Demonstrations*.

Torrent, T. T., Ellsworth, M., Baker, C. F., and Matos, E. E. d. S. (2018). The Multilingual FrameNet shared annotation task: A preliminary report. In *Proceedings of the International FameNet Workshop 2018: Multilingual FrameNets and Constructions*.

Torrent, T. T., Matos, E. E. d. L., Lage, L., Laviola, A., Tavares, T., Almeida, V., and Sigiliano, N. (Forthcoming). Towards continuity between the lexicon and the constructicon in FrameNet Brasil. John Benjamins.

Virk, S. M. and Prasad, K. V. S. (2018). Towards Hindi/Urdu framenets via the Multilingual Framenet. In *Proceedings of the International FameNet Workshop 2018: Multilingual FrameNets and Constructions*, pages 66–71.

Vossen, P., Fokkens, A., Maks, I., and van Son, C. (2018). Towards an open Dutch FrameNet lexicon and corpus. In *Proceedings of the International FameNet Workshop 2018: Multilingual FrameNets and Constructions*.

You, L. and Liu, K. (2005). Building Chinese FrameNet database. *Natural Language Processing and Knowledge Engineering, 2005. IEEE NLP-KE '05. Proceedings of 2005 IEEE International Conference on*, pages 301–306, Oct.-1 Nov.

Zawisławska, M., Derwojedowa, M., and Linde-Usiekniewicz, J. (2008). A FrameNet for Polish. In *Converging Evidence: Proceedings to the Third International Conference of the German Cognitive Linguistics Association (GCLA'08)*, pages 116–117.