Exploiting Frame Semantics and Frame-Semantic Parsing for Automatic Extraction of Typological Information from Descriptive Grammars of Natural Languages

Shafqat Mumtaz Virk[1], Azam Sheikh Muhammad[2], Lars Borin[3], Muhammad Irfan Aslam[4], Saania Iqbal[5], and Nazia Khurram[6]

[1,3]Språkbanken, University of Gothenburg, Sweden *{shafqat.virk, lars.borin}@svenska.gu.se*
[2]Department of CS & Eng., Chalmers University of Technology, Sweden, *azams@chalmers.se*
[4]University of Skövde, Sweden, *irfan.aslam@hotmail.com*
[5,6]GIFT University, Pakistan, *sania__iqbal@hotmail.com, nazia.yousaf@gmail.com*

Abstract

We describe a novel system for automatic extraction of typological linguistic information from descriptive grammars of natural languages, applying the theory of frame semantics in the form of frame-semantic parsing. The current proof-of-concept system covers a few selected linguistic features, but the methodology is general and can be extended not only to other typological features but also to descriptive grammars written in languages other than English. Such a system is expected to be a useful assistance for automatic curation of typological databases which otherwise are built manually, a very labor and time consuming as well as cognitively taxing enterprise.

1 Introduction

There are more than 7,000 living languages in the world and grammatical descriptions[1] are available for some 4,000 of these (Seifart et al., 2017). A central concern of the academic discipline of linguistics is to classify languages along different dimensions. The subbranch of linguistics which deals with classification and comparison of languages based on their structural and functional characteristics is known as *linguistic typology* (or *typological linguistics*. In addition to comparing the world's languages, practitioners of linguistic typology aim to explore the distribution of various structural and functional patterns among languages and to explain them in historical and/or universal terms. (Song, 2010)

To achieve its goals, typological linguistics has relied largely on manual reading of available descriptive material about languages for the extraction of pertinent feature values for comparison of these across languages. For example, in their sentence structure, different languages favor different word orders (e.g. subject-object-verb, verb-object-subject, or object-verb-subject, etc). If word order is to be used as one structural feature for language comparison, the word order of all the languages to be compared has to be found out manually by reading available material on those languages. This is doable, if the scope of comparison is to be limited to a few features spanning across a few languages. If one aims to extend the scope from a few to a few hundred features spanning across thousands of languages, the manual reading and comparison strategy seems simply unfeasible. With the availability of large amounts of digital data, and the recent advances in natural language processing, automatic extraction of typological features from grammatical descriptions seems a plausible task.

There have already been a few attempts to automatically extract typological and other linguistic information from descriptive grammars (Virk et al., 2017; Borin et al., 2018). In these studies, the authors have reported simple pattern matching and syntactic parsing based approaches, and have shown that their strategy is useful, yielding reasonable precision and recall values. Even though simple pattern matching based systems are easy to comprehend, implement, and maintain, they require a deep understanding of the rules/patterns which may not be very obvious in certain cases. Further, such systems are heuristics based and also require a large manual effort (Chiticariu et al., 2013). For these reasons the pattern-matching based systems are becoming a less and

[1] Grammatical descriptions are plain text descriptions of various phonological, morphological, and syntactic characteristics of languages.

Proceedings of Recent Advances in Natural Language Processing, pages 1247–1256,
Varna, Bulgaria, Sep 2–4, 2019.
https://doi.org/10.26615/978-954-452-056-4_143

less attractive choice and machine learning and big-data based approaches are taking their place.

In this paper, we report a novel methodology and a system for automatic extraction of typological information from descriptive grammars. The system is based on the theory of frame semantics and frame-semantic parsing, and employs a machine learning based approach. Using a set of domain-specific semantic frames, a handful of descriptive grammars are manually annotated with linguistic frames and their associated frame elements. Using these annotations as training data, machine learning models are trained and tested, which are then used to automatically annotate new descriptive grammars. The annotations are subsequently converted into typological feature values using a small rule based module, hence resulting in automatic extraction of typological feature values from descriptive grammars.

Section 2 describes frame semantics, FrameNet, and frame-semantic parsing as a theoretical background, followed by a brief introduction to the the linguistic domain FrameNet (Section 3). The development of a parser for the linguistic domain is outlined in Section 4, leading to a description of the system for automatic extraction of typological features (Section 5).

2 Frame Semantics, FrameNet, and Frame-Semantic Parsing

2.1 Frame Semantics

Frame semantics is a theory of meaning in language introduced by Charles Filmore (Fillmore, 1976, 1977, 1982). The theory is based on the notion that meanings of words can be best understood when studied in connection with the situations to which they belong, and/or in which they may occur.

The backbone of the theory is a conceptual structure called a *semantic frame*, which is a script-like description of a prototypical situation, an event, an object, or a relation. As an example, consider a real life scenario of robbery – a situation in which someone (a perpetrator) wrongs a victim by taking something (goods) from him/her. A structured representation of such a situation is called a *semantic frame*. The participants of the situation (i.e. the perpetrator, the victim, the goods, time, place, manner) are called *frame elements*. Some of the them (the perpetrator, the victim, and the goods) are necessary for the situation to make sense and are called *core frame elements*. Others like the place where the robbery took place, the manner in which it took place are called *non-core frame elements* (see Ruppenhofer et al., 2016 for details). Now, with the availability of a structured representation of the robbery situation, words like *hold up, mug, ransack, rifle, rob, stick up* can be better understood and analyzed.

2.2 FrameNet

The development of a lexico-semantic resource – FrameNet (Baker et al., 1998) – based on the theory of frame semantics was initiated in 1998 for English. In this lexical resource, generally referred to as simply FrameNet or Berkeley FrameNet (BFN), each of the semantic frames has a set of associated words (or *triggers*) which can evoke that particular semantic frame. The linguistic expressions for participants, props, and other characteristic elements of the situations (called *frame elements*) are also identified for each frame. In addition, each semantic frame is accompanied by example sentences taken from naturally occurring natural language text, annotated with triggers, frame elements and other linguistic information.

In the context of deploying FrameNets in NLP applications, BFN and other FrameNets have often been criticized for their limited coverage. A solution to this problem is to develop domain-specific (sublanguage) FrameNets to complement the corresponding general-language FrameNets for particular NLP tasks. In the literature we find such initiatives covering various domains, e.g.: (1) a FrameNet to cover medical terminology (Borin et al., 2007); (2) *Kicktionary*,[2] a soccer language FrameNet; (3) the *Copa 2014* project, covering the domains of soccer, tourism and the World Cup in Brazilian Portuguese, English and Spanish (Torrent et al., 2014).

Because of their perceived usefulness for a variety of purposes, general-language FrameNets have also been developed for a number of other languages including Chinese, French, German, Hebrew, Korean, Italian, Japanese, Portuguese, Spanish, and Swedish.

2.3 Frame-Semantic Parsing

In addition to the annotated examples, FrameNets are also often accompanied by varying amounts of frame-annotated natural running text intended

[2] http://www.kicktionary.de/

both to illustrate particular semantic-frame usages and to demonstrate the utility of frame semantics as a model of meaning in language. One of the uses of such annotated text is to develop automatic frame-semantic parsers, which in turn have proved useful in a number of natural language processing tasks including question answering (Shen and Lapata, 2007), coreference resolution (Ponzetto and Strube, 2006), paraphrase extraction (Hasegawa et al., 2011), machine translation (Wu and Fung, 2009), and information extraction (Surdeanu et al., 2003).

Frame-semantic parsing necessarily involves three basic steps. These are frame identification, frame element identification, and frame-element classification. Consider the annotated sentence shown in Figure 1 to better understand those basic steps of the frame semantic parsing. If the annotation shown is to be done automatically, the first step would be to consider each word of the sentence and check if it evokes a particular frame or not, and disambiguate in case if the candidate word evokes more than one frame. As a result, the word *agrees* (shown in bold) will be recognized as a lexical unit triggering the AGREEMENT frame. This is the frame identification task. Having identified the frame-triggering lexical units and the triggered frame, the next steps are to identify the text segments filling various semantic roles (i.e. frame elements) of the triggered frame. For this task, each word (or combination of words) in the sentence needs to checked for whether it expresses a frame element or not. This is the frame-element identification task. When a particular word or word-combination has been recognized as a frame element, it should be labeled next i.e. frame-element classification. So the frame-element identification and classification tasks will label the text segment *The genitive* as 'Participant_1', *sometimes* as 'Frequency', *noun* as 'Participant_2', *in gender* as 'Grammatical_Category' and *Gondi* as the 'Reference_Language'.

All of these three steps can be formulated as supervised machine learning classification tasks. Gildea and Jurafsky (2002) were the first to report their experiments with an automatic frame-semantic parsing system, and since then there have been a number of studies (Johansson and Nugues, 2008; Swayamdipta et al., 2017; Kabbach et al., 2018) and a shared task (Surdeanu et al., 2008) devoted to exploring and improving the task of frame-semantic parsing.

3 LingFN – a FrameNet for the Linguistic Domain

Linguistics has established a rich set of domain-specific terms and concepts such as *verbs*, *nouns*, *determiners*, *inflection*, *agreement*, *affixation*, etc. Inspired by other domain-specific FrameNets (mentioned in Section 2.2), the development of a FrameNet for the linguistic domain (LingFN) has been previously reported (Malm et al., 2018). LingFN contains two types of frames: the *filler frames* and the *eventful frames*. The former are to cover simple linguistic terms such as *noun*, *verb*, etc. mostly referring to the morpho-syntactic linguistic categories, and the later type covers linguistic processes such as *inflection*, *agreement*, *affixation* etc. Based on the empirical investigations of the usage of those terms and concepts in a large collection of domain-specific data, both types of frames were constructed. Consider again Figure 1 which also shows the structure of the AGREEMENT frame – an eventful of frame. In the linguistic domain, agreement is a phenomenon in which words of a particular morphological category (e.g. nouns) agree with another morphological category for a particular grammatical category (gender, number, etc.). The structure shown in Figure 1 was developed based on the investigations of the usage of the word *agree* within the linguistic corpora. See Malm et al. (2018) for a detailed description of the procedure followed to design frames together with annotated example sentences given to show the realization of those frames in the linguistic domain data.

The current version of LingFN contains a total of 100 frames, 32 frame elements, 360 lexical units, and around 2,800 annotated examples.

For the study reported in this paper, we have restricted ourselves to the frames outlined in Table 1 in addition to the AGREEMENT frame above. These frames will prove useful while automatically extracting values of certain typological feature as will be elaborated in Section 5.

4 The LingFN Parser

This section describes the development of an automatic parser based on LingFN. Treating it as a

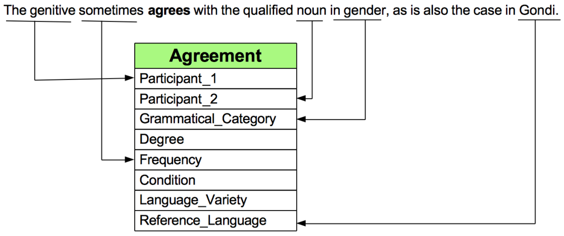


Figure 1: The structure of the AGREEMENT frame

supervised machine learning task, we describe the production of training data, feature selection and training data generation, and the training and testing of machine learning models in the following subsections.

4.1 Data Annotation

A set of 66 grammatical descriptions from the classical *Linguistic Survey of India* (LSI; Grierson, 1903–1927)[3] were annotated with the frames from LingFN described in Section 3.[4]

An online annotation tool from the Brazilian FrameNet Brasil project[5] was used. The annotation process was a collective effort and a number of data annotators were involved in this step. Each data annotator was responsible for 6 documents, and the length of each document is between 90 and 255 sentences. The task for each annotator was to go through the sentences of each document, identify each lexical unit and record the frame triggered by it. Once the frame has been identified, the next task was to identify and label the text segments (if present) of the sentence indicating the frame elements of the frame. Figure 2 shows a screenshot from the web tool used for annotation. As can be seen, the tool provides a layered view with the sentence to be annotated appearing in the top layer. This is followed by two layers, one for the frame and the other for frame elements annotation, for each of the triggered frames. The screenshot shows the annotation of the sentence *the verb agrees in number and gender with the subject* with two frames i.e. the frame VERBAL triggered by the lexical units *verb* (shown in black) with an empty FE layer (since there is no FE to be annotated in the sentence for this frame) and the frame AGREEMENT triggered by the lexical unit *agrees*. The FE layer for the AGREEMENT frame contains the annotations 'Participant_1' (in red) referring to text-segment *the verb*, 'Participant_2' (in blue) referring to the text segment *the subject*, and 'Grammatical_Category' (in green) referring to the text segment *in number and gender*.

1	**Frame Name:** AFFIXATION
	Definition: A frame to capture the phenomena of affixation in linguistics, which is the process of adding a morpheme — or affix— to a word to create either a different form of that word or a new word with a different meaning
	Frame Elements: Stem, Affix, Language_Variety, Reference_Language, Location Frequency, Manner, Purpose, Condition, Position, Degree
	Example Annotation: [An n]$_{\text{Morpheme\_one}}$ is [often]$_{\text{Frequency}}$ [infixed]$_{\text{LU}}$ [after the first vowel of a word]$_{\text{Morphosyntactic\_position}}$, the vowel being also repeated after n .
2	**Frame Name:** SEQUENCE
	Definition: A frame to capture the ordering information of various morphological or syntactical categories
	Frame Elements: Entity_1, Entity_2, Order, Language_Variety, Entities, Condition, Frequency, Reference_Language, Certainty, Data, Data_Translation
	Example Annotation: [Adjectives]$_{\text{Entity\_1}}$ in [Garo]$_{\text{Language\_Variety}}$, [as in Kachari]$_{\text{Reference\_Language}}$, [generally]$_{\text{Frequency}}$ [[follow]$_{\text{Order}}$]$_{\text{LU}}$ [the noun they qualify]$_{\text{Entity\_2}}$
3	**Frame Name:** CREATION
	Definition: A frame to capture the phenomena of creation of a morphological or syntactic category from another morphological or syntactic category
	Frame Elements: Created_Entity, Created_From, Process, Degree, Certainty , Language_Variety, Reference_Language, Data, Data_Translation, Condition
	Example Annotation: [Adverbs]$_{\text{formed\_Entity}}$ are often separate words , but are also [frequently]$_{\text{Degree}}$ [formed]$_{\text{LU}}$ from [the corresponding adjective]$_{\text{formed\_From}}$ [by adding hui or ui]$_{\text{Process}}$.

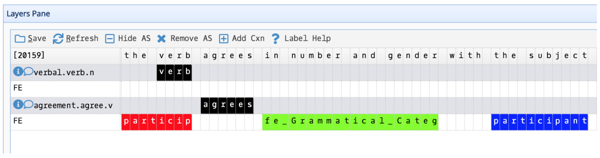
Table 1: Targeted frames

Table 2 shows some statistics of the produced annotated data. Further details, such as inter-annotator agreement, etc., are beyond the scope of

[3] The LSI presents a comprehensive survey of the languages spoken in South Asia conducted in the late nineteenth and the early twentieth century by the British government. It has descriptions of various phonological, morphological, and grammatical features of about 723 linguistic varieties spoken in the nineteenth-century British-controlled India (modern Pakistan, India, Bangladesh, and parts of Burma).

[4] There, we also mentioned that we have restricted ourselves to a few frames for the study reported in this paper, but it is worth mentioning that the data was annotated with the full set of LingFN frames.

[5] http://www.ufjf.br/framenetbr-eng/



[20159] the verb agrees in number and gender with the subject

verbal.verb.n: verb

FE

agreement.agree.v: agrees

FE: particip | fe_Grammatical_Categ | participant

Figure 2: Frame annotations

# Documents	# Sentences	# Frames	# Frame Elements
66	3,926	7,080	4,599

Table 2: Annotated data statistics

this paper, and will be reported on in a separate publication.

4.2 Feature Selection and Training Data Generation

To train machine learning based classification models, the first task is to define a useful set of features, and then compute the feature values from the training data. The area of frame-semantic parsing is well researched meaning that a set of suitable features both for the frame-element identification and frame-element classification tasks have previously been explored (Johansson and Nugues, 2008; Das et al., 2014). Since our objective in this work is not to improve frame-semantic parsing, but rather to show how frame-semantic parsing can be exploited to extract linguistic features from descriptive grammars, we have opted to use the same feature set as described by Johansson and Nugues (2008).

While a detailed explanation of the features can be found in Johansson and Nugues (2008), Table 5 lists *15* features used for training both the frame-element identification and frame-element classification models. The procedure for generating the training instances and computing the features values is as follows: Each sentence of the training data set was parsed using the Stanford constituency parser (Manning et al., 2014) resulting into parse trees as shown in Figure 3. Each node of the tree is then taken as one training instance and the required feature values are computed. The features values given in the last column of Table 5 were computed for the NP node referring to *the qualified nouns* (the one enclosed within the dotted area) as the argument node (i.e. the frame-element node) and with *agree* as the target word (i.e. frame triggering word). When computing for the whole tree, if a given argument node has been annotated as a frame element in the annotation the computed feature vector will get 'Y' as its class label, and 'N' otherwise resulting into the type of training instances shown in Table 3, and making it a binary classification task.

For the frame-element classification task, the objective is not to learn whether an argument node is a frame element or not, but rather to learn the frame-element label for all the annotated nodes (a multi-class classification task). The training data for the frame-element classification task was generated by going through all the nodes in the parse tree (as above), but this time only keeping those nodes which have been annotated as frame elements together with their label. Table 4 below shows a few instances from the generated frame-element classification data set.

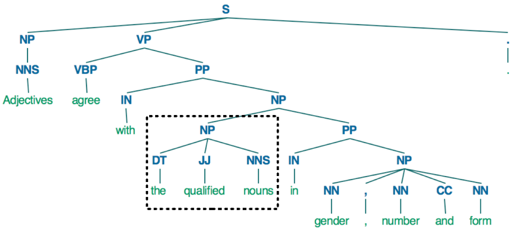


Figure 3: Example parse tree

The described procedure resulted into a set of 197,313 training instances for the frame-element identification task, and 11,904 training instances for the frame-element classification task. After removal of duplicates, 81,878 instances were left. Out of these, 76,036 (92.86%) were labeled 'Y', and the remaining 5,842 (7.14%) were labeled 'N'.

After removing duplicates, 5,855 cases were available for the frame-element classification task, covering 49 different classes of frame elements.

For the frame identification task, a simple dictionary lookup based approach was preferred at this stage simply because there are not many frames in the LingFN, indicating that frame disambiguation is rarely required. In future, we intend to train model for this task as well.

4.3 Data Representation

All of the variables in both of the datasets are the same, except that they differ on the possible set of values for the target variable, *label*. However, in either case, together with the target, all of the variables are categorical.

target_lemma	target_pos	arg_word	arg_word_pos	right_word	right_word_pos	left_word	left_word_pos	parent_word	parent_word_pos	c_subcat	phrase_type	position	fes_list	gov_cat	label
verb	NNS	also	RB	Default	Default	about	RB	walked	VBD	NP->JJNNS	ADVP	R	fe_language_variety#and#fe_data	VP	N
plural	NN	is	VBZ	Default	Default	was	VBD	past	NN	NP->NNNN	SBAR	L	fe_subclass#and#fe_data#and#fe_data_translation	ROOT	Y
plural	NN	past	NN	.	.	The	DT	ROOT	ROOT	NP->NNNN	ROOT	O	fe_subclass#and#fe_data#and#fe_data_translation	ROOT	N
oblique	JJ	ag	NN	Default	Default	twai-na	NN	twai	NN	ADJP->JJJJ	NP	R	fe_sublass#and#fe_data#and#fe_data_translation	VP	Y
decline	VBD	like	IN	Default	Default	Default	Default	like	IN	VP->VBDPP	IN	R	fe_inflectional_scheme#and#fe_form	VP	N

Table 3: A sample from the frame-element identification dataset

target_lemma	target_pos	arg_word	arg_word_pos	right_word	right_word_pos	left_word	left_word_pos	parent_word	parent_word_pos	c_subcat	phrase_type	position	fes_list	gov_cat	label
verb	VB	tong	NN	Default	Default	Default	Default	tong	NN	VP->VBSBAR	NN	R	fe_data#and#fe_data_translation#and#fe_subclass	VP	data
pronoun	NNS	Relative	JJ	Default	Default	Default	Default	pronouns	NNS	NP->DTJJNNS	JJ	L	fe_language_variety#and#fe_data	VP	sublass
prefix	NNS	towards	IN	Default	Default	signifying	VBG	Hon	NNP	NP->DTVBGNNS	UCP	R	fe_subclass#and#fe_data#and#fe_language_variety	VP	fe_Data_Translation
suffix	NN	yo	NN	Default	Default	Default	Default	ya	NN	NP->DTNN	NP	L	fe_subclass#and#fe_language_variety	VP	data
pronoun	NN	who	WP	Default	Default	Default	Default	who	WP	NP->JJNNNN	WP	R	fe_language_variety#and#fe_data#and#fe_data_tr	ROOT	data_translation

Table 4: A sample from the frame-element classification dataset

#	Feature	Explanation	Example Feature Value
1	target_lemma	Lemmatized form of the target word	agree
2	target_pos	Part of speech (POS) tag of the target_lemma	VBP
3	arg_word	The head word of the argument node	nouns
4	arg_word_pos	POS tag of the arg_word	NNS
5	right_word	The right most dependent word of the argument node	the
6	right_word_pos	POS tag of the right_word	DT
7	left_word	The left most dependent word of the argument node	NA
8	left_word_pos	POS tag of the left_word	NA
9	parent_word	Head word of the parent node of the target	agree
10	parent_word_pos	POS tag of the parent_word	VBP
11	c_subcat	Subcategorization frame corresponding to the phrase structure rule used to expand the phrase around the target	VP- >VBP PP
12	phrase_type	Phrase type of the argument node	NP
13	position	Position of the argument w.r.t target word	
14	fes_list	List of frame elements of the triggered frame	(Participant_1, Participant_2, Grammatical_Category, Degree, Frequency, Language_Variety, Reference_Language, Condition)
15	gov_cat	The governing category either S or VP	VP

Table 5: Feature set

In order to achieve best performance while performing machine learning modeling, the right choice of data representation technique for categorical data is very important. The main reason is that there are a limited number of machine learning algorithms that can be directly applied to categorical data. On the other hand, if we can turn them into numerical variables, starting from basic Decision Trees, Naïve Bayes, Support Vector Machines, Logistic Regression, Random Forest, to Multi-layer Perceptron (Deep Learning), almost all of the machine learning algorithms can be applied. There are plenty of techniques to transform categorical values to numerical data. One such technique is one-hot encoding. The basic strategy is to convert each category level (value) of the categorical variable into a new variable, and assign the value *1* to this new variable wherever the corresponding categorical variable equals this level, and *0* otherwise. This is done for all category levels of the variable being encoded except one, which will be redundant (applies when all other associated variables equal zero) and can be any category level. The key is to always create one fewer binary variables than the number of categories. The new binary variables together replace the original categorical variable. The new variables are sometimes termed *dummy variables*, and the approach is also called *Dummy Variables Encoding*. This encoding has the benefit of not weighting a value improperly, but does have the downside of adding more variables to the dataset.

4.4 Model Training

Successful encoding makes the dataset ready to be used for applying machine learning algorithms. We experimented with different machine learning models. A comparison of the machine learning models chosen for binary classification (frame-element identification) and multiclass classification (frame-element classification) tasks for our datasets has been performed. Tables 6 and 7 provide a comparison of various evaluation metrics using average scores of 5-fold cross validation) respectively for the frame-element identification and classification tasks.

Model	Accuracy	Precision	Recall	F_score
Decision Tree	0.926	0.712	0.664	0.921
Logistic Regression	0.936	0.797	0.609	0.922
Naïve Bayes	0.658	0.552	0.686	0.741
Support Vector Machine	0.929	0.465	0.5	0.895

Table 6: Model comparison for the frame-element identification dataset using one hot encoding

Model	Accuracy	Precision	Recall	F_score
Decision Tree	0.789	0.56	0.542	0.786
Logistic Regression	0.817	0.619	0.545	0.808
Naïve Bayes	0.528	0.481	0.489	0.537
Support Vector Machine	0.465	0.019	0.04	0.295

Table 7: Model comparison for the frame-element classification dataset using one hot encoding

The best performing logistic regression models were selected and used in the typological feature extraction system described in the next section.

5 Topological Feature Extraction System

Figure 4 shows the complete architecture of the typological feature extraction system. As shown (the middle part within dotted area), the system takes a descriptive grammar in raw form and annotate it with LingFN frames using the pre-trained models both for the frame-element identification and frame-element classification tasks (i.e. the part above the dotted area). The annotated data is further processed with a simple rule based module to convert those annotations to typological feature values (i.e. the part below the dotted area). Lets take an example to explain this part in particular, and the overall purpose of such a system in general.

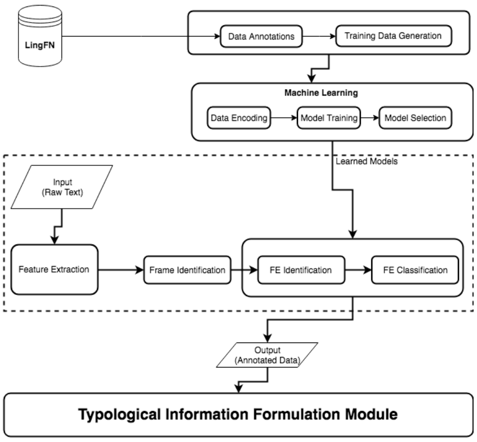
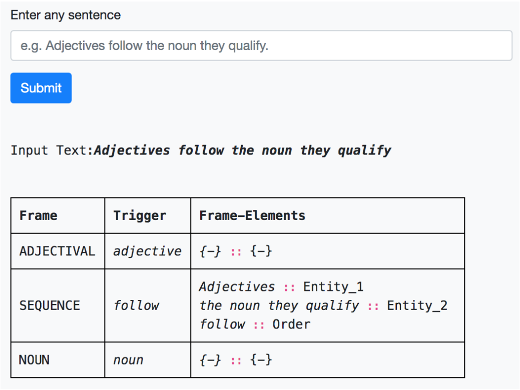


Figure 4: System architecture

Suppose we are interested in finding an answer to the question "What is the order of adjective and noun in the noun phrase" for the Siyin[6] language. The LSI data set contains a grammatical description of this language, and one of the sentences in that description is *The adjectives follow the noun they qualify*. Automatic parsing of this sentence using the developed LingFN parser will result into the annotations shown in Figure 5 (a screenshot from the web demo of the parser).



Input Text:***Adjectives follow the noun they qualify***

Frame	Trigger	Frame–Elements
ADJECTIVAL	*adjective*	{-} :: {-}
SEQUENCE	*follow*	*Adjectives* :: Entity_1 *the noun they qualify* :: Entity_2 *follow* :: Order
NOUN	*noun*	{-} :: {-}

Figure 5: Automatic frame annotation

This parse contains the answer to the above asked question. However, the typological databases often record answers in a specific format. For example, the answer to the above question could be required to be of one of these values 'NA', 'AN', or 'Both' meaning that the order is 'Adjective-Noun', 'Noun-Adjective', or 'Both' respectively. If required, the above given parse information can be converted into specific feature values using a simple rule-based module such as given below (only a part of the full module is shown). The module simply checks the contents of different frame elements to formulate the feature value.

Using the same sort of procedure and the frames mentioned in Section 3, we have targeted to extract and formulate values for some of the typological features given in the Grambank[7] and other typological databases. A few of these features are given below.

- Can an adnominal property word agree with the noun in gender/noun class?
- Can an article agree with the noun in gender/noun class?
- Can an article agree with the noun in number?
- Can the relative clause precede the noun?

[6] A Tibeto-Burman language spoken in southern Tedim township, Chin State, Burma. Also known as Siyin Chin and Sizang Chin, ISO 639-3: csy

[7] A typological database: https://github.com/clld/grambank.

Algorithm 1 Extract adjective noun order

```
procedure ExtractAdjective-NounOrder($parse$)
    for <every frame in parse> do
        if $frame = SEQUENCE$ then
            $NA \leftarrow False$
            $AN \leftarrow False$
            $Both \leftarrow False$
            if $'adjective' \in Entity\_1 \wedge' noun' \in Entity\_2$ then
                if $Frequency \in [sometimes, usually, mostly, often]$ then
                    $Both \leftarrow True$
                else if $order = follow$ then
                    $AN \leftarrow True$
                else if $order = precede$ then
                    $NA \leftarrow True$
                end if
            end if
        end if
    end for
end procedure
```

- Can the relative clause follow the noun?
- Order of Adjective and Noun.
- Order of Subject, Object and Verb.
- Order of Numeral and Noun.
- Order of Relative Clause and Noun.

It is worth mentioning that the same methodology can be used to extract values for various other typological features from the descriptive grammars. This will require designing suitable frames, annotating the data and re-training models. Further, the methodology can be extended to descriptive grammars written in languages other than English.

6 Conclusions and Future Work

We have presented a novel system for automatic extraction of typological features from descriptive grammars based on the theory of frame semantics and frame-semantic parsing. We have presented the methodology, set up the machinery and architecture, and shown the working of this machinery for extraction of feature values of an example typological feature. The methodology is scalable and can easily be extended not only to other features but also to the descriptive grammars written in other natural languages. This is required because there are many grammatical descriptions written in languages other than English (German, French, Spanish, and Russian are among them).

The system we report is expected to be a useful assistance for the development of typological databases, which otherwise are built manually. Manual curation of typological databases is very time and labor consuming, as well as cognitively taxing, thus making the scope of studies based on such databases very limited. We hope with the automatic extraction of typological databases, the scope of studies in typological and other related areas can be broaden further.

The current version of LingFN provides a very limited number of eventful frames restricting us to target only a few typological features. There are 195 typological features listed in Grambank. In the future, we would like to build more frames, annotate more grammars, and automatically extract values for as many as possible features of the Grambank.

In conclusion, the current study can be considered as a proof of concept. In the future, we plan to extend the system and evaluate it against existing manually curated typological databases to compute measures such as precision and recall. Further, the extraction of typological features is just a case study, the automatically annotated grammars are envisioned to be equally useful in other linguistic subdisciplines, in particular the related areas of genetic and areal linguistics. In the future, we also have plans to show the usefulness of the annotated descriptions in these and other related areas.

Acknowledgments

The work presented here was funded partially by the Swedish Research Council as part of the project *South Asia as a linguistic area? Exploring big-data methods in areal and genetic linguistics* (2015–2019, contract no. 421-2014-969), and partially by the *Dictionary/Grammar Reading Machine: Computational Tools for Accessing the World's Linguistic Heritage* (DReaM) Project awarded 2018–2020 by the Joint Programming Initiative in Cultural Heritage and Global Change, Digital Heritage and Riksantikvarieämbetet, Sweden.

References

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of ACL/COLING 1998*. ACL, Montreal, pages 86–90. https://doi.org/10.3115/980845.980860.

Lars Borin, Maria Toporowska Gronostaj, and Dimitrios Kokkinakis. 2007. Medical frames as target and tool. In *FRAME 2007: Building Frame Semantics Resources for Scandinavian and Baltic Languages. (Nodalida 2007 Workshop Proceedings)*. NEALT, Tartu, pages 11–18.

Lars Borin, Shafqat Mumtaz Virk, and Anju Saxena. 2018. Language technology for digital linguistics: Turning the Linguistic Survey of India into a rich source of linguistic information. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*. Springer, Cham, pages 550–563.

Laura Chiticariu, Yunyao Li, and Frederick Reiss. 2013. Rule-based information extraction is dead! Long live rule-based information extraction systems! In *Proceedings of EMNLP 2013*. ACL, Seattle, pages 827–832.

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

Charles J. Fillmore. 1976. Frame semantics and the nature of language. *Annals of the New York Academy of Sciences* 280(1):20–32. https://doi.org/10.1111/j.1749-6632.1976.tb25467.x.

Charles J. Fillmore. 1977. Scenes-and-frames semantics. In Antonio Zampolli, editor, *Linguistic Structures Processing*, North Holland, Amsterdam, pages 55–81.

Charles J. Fillmore. 1982. Frame semantics. In Linguistic Society of Korea, editor, *Linguistics in the Morning Calm*, Hanshin Publishing Co., Seoul, pages 111–137.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics* 28(3):245–288. https://doi.org/10.1162/089120102760275983.

George A. Grierson. 1903–1927. *A Linguistic Survey of India*, volume I–XI. Government of India, Central Publication Branch, Calcutta.

Yoko Hasegawa, Russell Lee-Goldman, Albert Kong, and Kimi Akita. 2011. FrameNet as a resource for paraphrase research. *Constructions and Frames* 3(1):104–127.

Richard Johansson and Pierre Nugues. 2008. The effect of syntactic representation on semantic role labeling. In *Proceedings of COLING 2008*. ACL, Manchester, pages 393–400.

Alexandre Kabbach, Corentin Ribeyre, and Aurélie Herbelot. 2018. Butterfly effects in frame semantic parsing: Impact of data processing on model ranking. In *Proceedings of COLING 2018*. ACL, Santa Fe, pages 3158–3169.

Per Malm, Shafqat Mumtaz Virk, Lars Borin, and Anju Saxena. 2018. LingFN : Towards a framenet for the linguistics domain. In *Proceedings of the IFNW 2018 Workshop on Multilingual FrameNets and Constructicons at LREC 2018*. ELRA, Miyazaki, pages 37–43.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of ACL 2014*. ACL, Baltimore, pages 55–60. http://www.aclweb.org/anthology/P/P14/P14-5010.

Simone Paolo Ponzetto and Michael Strube. 2006. Exploiting semantic role labeling, WordNet and Wikipedia for coreference resolution. In *Proceedings of HLT 2006*. ACL, New York, pages 192–199. http://www.aclweb.org/anthology/N/N06/N06-1025.

Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, Christopher R. Johnson, Collin F. Baker, and Jan Scheffczyk. 2016. *FrameNet II: Extended Theory and Practice*. ICSI, Berkeley.

Frank Seifart, Nicholas Evans, Harald Hammarström, and Stephen C. Levinson. 2017. Language documentation twenty-five years on. *Language* 94(4):e324–e345.

Dan Shen and Mirella Lapata. 2007. Using semantic roles to improve question answering. In *Proceedings of EMNLP-CoNLL 2007*. ACL, Prague, pages 12–21. http://www.aclweb.org/anthology/D/D07/D07-1002.

Jae Jung Song, editor. 2010. *The Oxford Handbook of Linguistic Typology*. Oxford University Press, Oxford.

Mihai Surdeanu, Sanda Harabagiu, John Williams, and Paul Aarseth. 2003. Using predicate-argument structures for information extraction. In *Proceedings of ACL 2003*. ACL, Sapporo, pages 8–15. http://www.aclweb.org/anthology/P03-1002.

Mihai Surdeanu, Richard Johansson, Adam Meyers, Lluís Màrquez, and Joakim Nivre. 2008. The CoNLL-2008 shared task on joint parsing of syntactic and semantic dependencies. In *Proceedings of CoNLL 2008*. ACL, Manchester, pages 159–177.

Swabha Swayamdipta, Sam Thomson, Chris Dyer, and Noah A. Smith. 2017. Frame-semantic parsing with softmax-margin segmental RNNs and a syntactic scaffold. *CoRR* abs/1706.09528.

Tiago Timponi Torrent, Maria Margarida Martins Salomão, Ely Edison da Silva Matos, Maucha Andrade Gamonal, Júlia Gonçalves, Bruno Pereira de Souza, Daniela Simões Gomes, and Simone Rodrigues Peron-Corrêa. 2014. Multilingual lexicographic annotation for domain-specific electronic dictionaries: The Copa 2014 FrameNet Brasil project. *Constructions and Frames* 6(1):73–91.

Shafqat Virk, Lars Borin, Anju Saxena, and Harald Hammarström. 2017. Automatic extraction of typological linguistic features from descriptive grammars. In *Proceedings of TSD 2017*. Springer, Cham, pages 111–119.

Dekai Wu and Pascale Fung. 2009. Semantic roles for SMT: A hybrid two-pass model. In *Proceedings of HLT-NAACL 2009*. ACL, Boulder, pages 13–16. http://dl.acm.org/citation.cfm?id=1620853.1620858.