LingFN: A Framenet for the Linguistic Domain

Shafqat Mumtaz Virk[1] (0000-0002-5030-9191), Per Malm[2] (0000-0002-6693-897X), Lars Borin[1] (0000-0001-5434-9329), Anju Saxena[3] (0000-0002-4736-3453)

[1]Språkbanken, University of Gothenburg; [2]Department of Scandinavian Languages, Uppsala University; [3]Department of Linguistics and Philology, Uppsala University
virk.shafqat@gmail.com, per.malm@nordiska.uu.se, lars.borin@svenska.gu.se, anju.saxena@lingfil.uu.se

Abstract. Frame semantics is a theory of meaning in natural language, which defines the structure of the lexical semantic resources known as framenets. Both framenets and frame semantics have proved useful for a number of natural language processing (NLP) tasks. However, in this connection framenets have often been criticized for their limited coverage. A proposed reasonable-effort solution to this problem is to develop domain-specific (sublanguage) framenets to complement the corresponding general-language framenets for particular NLP tasks, and in the literature we find such initiatives covering domains such as medicine, soccer, and tourism. In this paper, we report on building a framenet to cover the terms and concepts encountered in descriptive linguistic grammars (written in English) i.e. a framenet for the linguistic domain (LingFN) to complement the general-language BFN.

1 General Background and Introduction

Frame semantics is a theory of meaning introduced by Charles J. Fillmore and colleagues [8, 9, 10]. The backbone of the theory is a conceptual structure called **a semantic frame**, which is a script-like description of a prototypical situation that may occur in the real world, and can refer to a concept, an event, an object, or a relation. The major idea is that word meanings are best understood when studied from the point of view of the situations to which they belong, rather than looking at them separately and analyzing them individually. For example, consider a situation in which someone (the perpetrator) carries and holds someone (the victim) against his/her will by force – a kidnapping situation. With reference to this situation, words like *kidnap*, *abduct*, *nab*, *snatch*, etc. are easier to understand, which also involves identification and linking of the participants of the situation (perpetrator, victim, purpose, place, etc in the case of the *kidnapping* situation). In the world of frame semantics, the words which evoke semantic frames are called *lexical units* (LU) or *triggers*, and the participants and props in those frames are known as frame elements (FE). There are two types of FEs: core and non-core. Frame elements which are obligatory for the situation to make sense are *core*, while others are optional, i.e. *non-core*. For example, in the case of a KIDNAPPING frame, the PERPETRATOR and VICTIM are core, while peripheral elements like PURPOSE, PLACE, MANNER, etc., are non-core FEs.[1]

[1] Labels of FEs and frames are conventionally set in small caps.

FrameNet [1] – also commonly known as Berkeley FrameNet (BFN) – is a lexical semantic resource for English which is based on the theory of frame semantics. BFN contains descriptions of real world situations (i.e. semantic frames) along with the participants of those situations. Each of the semantic frames has a set of associated words (i.e. LUs) which evoke a particular semantic frame. The participants of the situations (i.e. FEs) are also identified for each frame. In addition, each semantic frame is coupled with example sentences taken from spoken and written discourse. FrameNet 1.5 has 1,230 semantic frames, 11,829 lexical units, and 173,018 example sentences. Further, a set of documents annotated with semantic frames and their participant information is also provided. In naturally occurring language, words mostly do not stand on their own, rather are interlinked through various syntactic, semantic, and discourse relations. To cope with this property of natural language, FrameNet has defined a set of frame-to-frame relations, and has proposed connections of certain frames to certain other frames thus providing a network of frames (hence the name FrameNet).

BFN and its annotated data have proved to be very useful for automatic shallow semantic parsing [11], which itself has applications in a number of natural language processing (NLP) tasks including but not limited to information extraction [18], question answering [17], coreference resolution [16], paraphrase extraction [13], and machine translation [21]. Because of their usefulness, framenets have also been developed for a number of other languages (Chinese, French, German, Hebrew, Korean, Italian, Japanese, Portuguese, Spanish, and Swedish), using the BFN model. This long standing effort has contributed extensively to the investigation of various semantic characteristics of many languages at individual levels, even though most crosslinguistic and universal aspects of the BFN model and its theoretical basis still remain to be explored.[2] Though framenets and the frame-annotated data have proved to be very useful both for the linguistic and the NLP communities, they have often been criticized for their lack of cross-linguistic applicability and limited coverage. In an ongoing project,[3] attempts are now being made to align framenets for many languages in order to investigate and test the cross-linguistic aspects of the FrameNet and its underlying theoretical basis. A proposed reasonable-effort solution to the coverage issue is to develop domain-specific (sublanguage) framenets to augment and extend the general-language framenet. In the literature, we can find such initiatives where domain-specific framenets are being developed, e.g.: (1) medical terminology [6]; (2) *Kicktionary*,[4] a soccer language framenet; and (3) the *Copa 2014* project, covering the domains of soccer, tourism and the World Cup in Brazilian Portuguese, English and Spanish [19].

Like many others, the area of linguistics has developed a rich set of domain specific terms and concepts (e.g. *inflection*, *agreement*, *affixation*, etc.). Such terms have been established by linguists over a period of centuries in the course of studying, inves-

[2] Most of the framenets – including BFN – have been developed in the context of linguistic lexicology, even if several of them have been used in NLP applications (again including BFN). The Swedish FrameNet (SweFN) forms a notable exception in this regard, having been built from the outset as a lexical resource for NLP use and only secondarily serving purposes of linguistic research [2, 4].

[3] http://www.ufjf.br/ifnw/

[4] `http://www.kicktionary.de/`

tigating, describing and recording various linguistic characteristics of a large number of different languages at phonological, morphological, syntactic, and semantic levels. For various computational purposes, attempts have been made to create inventories of such terms and keep a record of them, e.g.: (1) the *GOLD*[5] ontology of linguistic terms; (2) the *SIL glossary of linguistic terms*;[6] (3) the *CLARIN concept registry*;[7] and (4) *OLiA* [7].

A minority of the terms in the collections above are used only in linguistics (e.g., *tense*), and in many cases, non-linguistic usages are either rare (e.g., *affixation*) or specific to some other domain(s) (e.g., *morphology*). Others are polysemous, having both domain-specific and general-language senses. For example, in their usage in linguistics the verb *agree* and the noun *agreement* refer to a particular linguistic (morphosyntactic) phenomenon, viz. where a syntactic constituent by necessity must reflect some grammatical feature(s) of another constituent in the same phrase or clause, as when adjectival modifiers agree in gender, number and case with their head noun. This is different from the general-language meaning of these words, implying that their existing FN description (if available) cannot be expected to cover their usage in linguistics, which we will see below is indeed the case. Naturally, we need to build new frames, identify their LUs and FEs, and find examples in order to cover them and make them part of the general framenet if we are to extend its coverage. This is one of the major objectives of the work we report in this paper. The other objective is to investigate the relational aspects of the resulting linguistic frames. In the GOLD ontology, attempts were made to divide and organize the linguistic concepts into various groups. This organization is not without problems. GOLD seems to lack a theoretical foundation, and the validity of the organization remains untested. Also there is only one type of default relation – IS-A – between the terms/concepts. We aim to extend the relational structure of linguistic frames by exploring new relation types between the linguistic terms/concepts and building a network (i.e. Linguistic FrameNet – LingFN) of them.

The rest of the paper is organized as follows: Section 2 briefly describes the data sets that we have used, Section 3 outlines the architecture of the framenet, and the methodology is given in Section 4. This is followed by the application of the linguistic framenet (Section 5) and the conclusion and future work (Section 6).

2 The Data

The *Linguistic Survey of India* (LSI) [12] presents a comprehensive survey of the languages spoken in South Asia. It was conducted in the late nineteenth and the early twentieth century by the British government, under the supervision of George A. Grierson. The survey resulted in a detailed report comprising 19 volumes of around 9,500 pages in total. The survey covered 723 linguistic varieties representing the major language families and some unclassified languages, of almost the whole of nineteenth-century British-controlled India (modern Pakistan, India, Bangladesh, and parts of Burma). Im-

[5] `http://linguistics-ontology.org/`
[6] `http://glossary.sil.org`
[7] `https://www.clarin.eu/ccr`

portantly for our purposes, for each major variety it provides a grammatical sketch (including a description of the sound system).

The LSI grammar sketches provide basic grammatical information about the languages in a fairly standardized format. The focus is on the sound system and the morphology (nominal number and case inflection, verbal tense, aspect, and argument indexing inflection, etc.), but there is also syntactic information to be found in them. Despite its age, it is the most comprehensive description available of South Asian languages, and since it serves as the main data source in a large linguistic project in which we are involved, it is natural for us to use it as a starting point for the development of LingFN, but in the future we plan to extend our range and use other publicly available digital descriptive grammars.

3 The General Architecture of LingFN

Figure 1 shows the general architecture of LingFN.

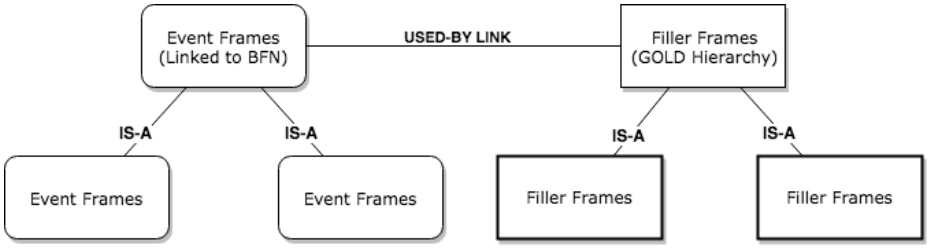

Fig. 1: The basic structure of LingFN

3.1 Frame Types

As shown in Figure 1, there are two main types of frames: *filler frames* and *event frames*. The frames which are comparatively a bit more complex in their structure and represent eventful types of scenes (or concepts) are called event frames. In contrast filler frames are simple in their structure and may act as slot fillers to the event frames. To explain these two types of frames further, let us take an example frame from the BFN. The BORROW frame holds information about a situation involving a BORROWER that takes possession of a THEME belonging to a LENDER under the tacit agreement that the THEME be returned after a DURATION of time. Observe the annotated sentence below:

(1) Does [$_{\text{BORROWER}}$ my Mum] [$_{\text{LU}}$ borrow] [$_{\text{THEME}}$ money] [$_{\text{LENDER}}$ off you]? (BFN)

The LU in example 1 evokes an entire scene containing various roles, or FEs. These type of complex frames are referred to as *event frames* in LingFN. A corresponding example from the linguistic domain could be the AGREEMENT frame.Consider the following annotated example:

(2) [The PARTICIPANT-1 participle] [LU agrees] in [GRAMMATICAL_CATEGORY gender and number] with [PARTICIPANT_2 the object] [CONDITION if the latter is in the form of the nominative].

These stand in contrast to frames with a less eventful structure (i.e. *filler frames*, which usually fill in the roles of the event frames). An annotated example of a filler frame (the MONEY frame) is given below in bold, followed by the PARTICIPLE frame from the linguistic domain.

(3) Does [BORROWER my Mum] [LU borrow] [THEME **[MONEY money]**] [LENDER off you]? (BFN)

(4) [The PARTICIPANT-1 **[PARTICIPLE participle]**] [LU agrees] in [GRAMMATICAL_CATEGORY gender and number] with [PARTICIPANT_2 the object] [CONDITION if the latter is in the form of the nominative].

3.2 Frame-to-Frame Relations

The above described frame types are ordered hierarchically by the (frame-to-frame) IS-A relation. This relation is intended to preserve the GOLD ontology structure by linking a top-level frame to a frame at the next lower level. For example to preserve the fact that a clitic is a morpheme, the corresponding CLITIC frame is linked to the MORPHEME frame by an IS-A relation. Traditionally, the IS-A relation is used for a much richer inheritance type of linkage, where the lower level entity inherits certain attributes from the upper level entity. At this stage we are using IS-A in a simpler sense to preserve only the structural level information but in future, we intend to enhance this type of linking.

As mentioned above, filler frames may appear as FEs of the event frames. Any observation of a filler frame occurring in an event frame is documented in LingFN by means of connecting the frames with a used-by link. For example, consider the annotated example sentence from the LINGUISTIC PLACING frame below.

(5) [FORM Adverbs] [COP are] [LU placed] [MORPHOSYNTACTIC_POSITION before adjectives and after verbs] . (LINGUISTIC_PLACING)

The FEs in example 5 contain some words that are relevant for the linguistic domain, such as, *adverb*, *adjective* and *verb*. These words evoke certain filler frames, e.g. GOLD_VERBAL, GOLD_ADJECTIVAL, and GOLD_PART_OF_SPEECH_PROPERTY. When an event frame FE contains an LU that is found in the filler frames, this connection is recorded in LingFN as used-by link.

4 Methodology

The development of a domain specific framenet involves (1) selection of a framenet development strategy; (2) identification and construction of of new domain specific frames; and (3) identification and development of FEs of the newly identified frames. In this section, we describe the methodology for each of these stages.

4.1 Framenet Development

At the framenet level, there are at least four different types of methodologies which have been previously discussed in the literature, namely *Lexicographic Frame-by-Frame*; *Corpus-Driven Lemma-by-Lemma*; *Full-Text*; and *Domain-by-Domain*.

In our case, we have used a hybrid of the lexicographic and the corpus-driven approach. The choice is largely driven by the available resources and the objectives of the project for which LingFN is being developed. As mentioned above, there are available inventories of linguistic terms and concepts. This means one could start with entries in one of those inventories (e.g. GOLD) and build semantic frames for them (i.e., the frame-by-frame strategy). This is actually how we started. As can be expected, GOLD's coverage is limited. Also, the objectives of the project require that we cover the LSI corpus. So, we develop new frames when encountering corpus data that is unattested in GOLD (i.e., the corpus-driven approach).

4.2 Frame Identification and Construction

Frame Identification Before we can construct a frame, we need to decide when and what domain-specific frames we need to design. In the first round, we started with the GOLD list of linguistic terms and developed corresponding frames. In the second round, we scanned through the LSI corpus, and began developing frames for the linguistic terms/concepts found in the corpus but not in GOLD. Here, we were faced with an additional issue of deciding which terms/concepts are specific to the linguistic domain and resolving the ambiguous cases. Since we are using a domain specific corpus, an assumption in this regard could be that the terms within a domain-specific corpus are mostly related to that particular domain. Since this can not be guaranteed, we have to deal with the polysemous occurrence of the terms. For this purpose, we have used *semi-automatic uniqueness differentiation* (SUDi) [14], a corpus-driven statistical method to judge the polysemous nature of the lemmas.

Frame Construction Once we have chosen a frame (in the case of the GOLD list), or identified a frame (in the case of scanning through the LSI corpus), to be developed, we construct the corresponding frame and put it into a frame repository. Constructing a frame requires three things: (1) identification of frame LUs; (2) identification of FEs; and (3) creation and storage of the frame as a lexical entry in a dedicated lexical database.

For lexical unit identification, we have largely relied on linguistic intuition and general linguistic knowledge. For example, for the semantic frame AFFIXATION, the identified list of lexical units is *{affix.v, affixed.a, infix.v, infixed.a, postfixed.a, prefix.v, suffix.v, suffixed.a}*. The FE identification involves the recognition of various semantic roles of the frame. We have used a corpus-driven approach for the identification purpose, which is described in detail in 4.3. For the third part, we have used a web-based frame editor provided by Språkbanken's Karp infrastructure [5]. The tool was built as part of the Swedish FrameNet++ project [3, 4], and was used for creation of Swedish frames. Figure 2 shows a snapshot of the editor, and as can be seen, the creation of a new frame means filling in various fields making part of the structure of the frame. Most of the

fields shown in the figure should be self-explanatory, but the following may require explanation:

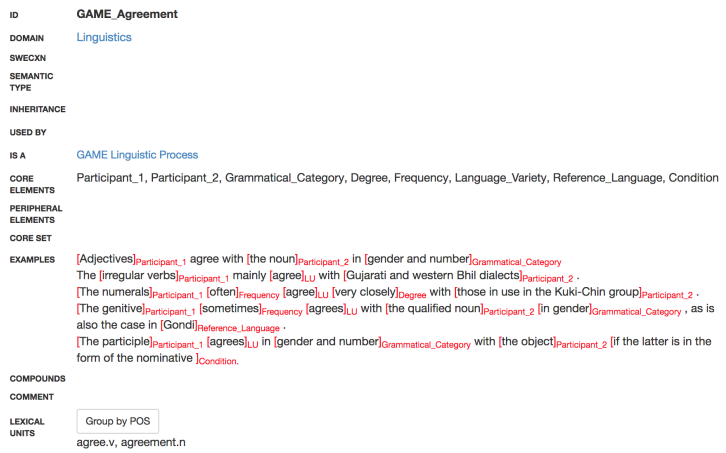

Fig. 2: A frame in the Karp editor

Used-by Links: This entry holds the names of the event frames that uses the filler frames. For instance, take the LINGUISTIC PLACING frame. This frame contains various verbal and adjectival words denoting acts of placing things. An annotated example follows:

(6) It will be seen that [THEME the personal pronoun which we translate as a possessive] is [FREQUENCY often] [LU put] [GRAMMATICAL_CATEGORY in the nominative] [GOAL before such prefixes] .

The following result could be seen, if the example above were annotated with the filler frames (fillers in boldface):

(7) It will be seen that [THEME the **[PRONOUN pronoun]** which we translate as a possessive] is [FREQUENCY often] [LU put] [GRAMMATICAL_CATEGORY in the **[CASE nominative]**] [GOAL before such **[AFFIX prefixes]**] .

So, from the example above, it is seen that the LUs from the following GOLD frames are used by the event frame:

- Frame: GOLD_PRONOUN = [SUBCLASS possessive] [SUBCLASS personal] [LU pronoun]
- Frame: GOLD_CASE = [LU nominative]
- Frame: GOLD_AFFIXATION = [LU prefixes]

The Core Elements and Peripheral Elements: These two fields are supposed to contain a list of core and non-core FEs belonging to the frame. At this stage we do not distinguish between these FE types, but in the future the plan is to make use of this distinction. The process of identifying the FEs for a given frame is described in Section 4.3. Below we list various FEs together with a brief description of the type of information they contain.

- Language_Variety: This FE is intended to record the name of a language variety.
- Reference_Language: Often, while describing a particular aspect of a language a reference is made to another language (e.g. in the sentence 'The verb agrees with its subject in gender , number and person as in Hindostani' Reference_Language is 'Hindostani' .)
- Data: Often, examples are given to describe a particular aspect, this FE is intended to record those examples.
- Data_Translation: The English glossing of examples recorded by the Data FE.
- Subclass: The subclass of the the entity/phenomenon being described by the frame e.g. 'interrogative' in the 'interrogative adjective'.
- Position: The position of the entity being described e.g. 'initial' in 'initial soft consonants....'
- Frequency: The frequency of the phenomenon being described e.g. 'often' in 'Adjectives are often followed by a suffix...'
- Degree: The degree of the phenomenon being described e.g. 'strong' in 'Initial soft consonants are pronounced with a very strong aspiration.'
- Condition: The condition in which the phenomenon being described occurs e.g. 'if the object is of the second person' in 'subject of the first person is not separately marked if the object is of the second person.'
- Means: The means by which the described phenomenon is materialized e.g. 'by means of postpositions' in 'The pronouns are inflected like nouns by means of postpositions '
- Manner: The manner in which the described phenomenon takes place e.g. 'without any suffix' in 'The genitive is apparently formed by prefixing the governed to the governing word without any suffix'.
- Certainty: The certainty of the phenomenon being described e.g. 'may' in 'Both the hard and the soft sounds aspirant may be either aspirated or unaspirated'.

Examples: A list of example sentences from the LSI corpus annotated with LUs and FEs is provided in this field. We have provided at least 20 annotated examples per frame. Some annotated sentences are provided below. The parentheses to the right contain the name of the frames where these peripheral FEs appear.

(8) a. [LANGUAGE_VARIETY Burmese] [LU orthography] (ORTHOGRAPHIC_SYSTEM)
b. [LU sentence] [DATA m'tā teku bri no], [DATA_TRANSLATION he rice to-buy wishes], [DATA_TRANSLATION he wants to buy rice] (ORTHOGRAPHIC_PHRASE)
c. [NATIONALITY Danish] [LU philologist] [NAME Rask] (SPECIALIZED_LINGUISTS)
d. [PLACE final] [LU vowels] (SEGMENT)
e. [AUTHOR Mr. Godwin Austen' s] [LU vocabulary] (LINGUISTIC_DATA_STRUCTURE)
f. [BRANCH comparative] [LU philology] (LINGUISTIC_SUBFIELD)

4.3 Frame Element Identification and Development

After we have realized the need for a new domain specific frame (i.e. after dealing with polysemy), we have to design the structure of the frame which involves identification

Head	Dependent	Sentence
		Relation: advcl (adverbial clause)
'agrees'	'if the latter is in the form of the nominative'	'The participle agrees in gender and number with the object if the latter is in the form of the nominative .'
'agrees'	'when mutable'	'in the case of Transitive verbs , the subject is put in the agent case , and , when mutable , the verb agrees in gender and number with the object .'
		Relation: advmod (adverbial modifier)
'agree'	'apparently'	'The personal pronouns apparently also agree .'
'agree'	'often'	'The numerals often agree very closely with those in use in the Kuki-Chin group .'
'agrees'	'closely'	'The Gondi of Mandla closely agrees with the preceding sketch .'
		Relation: nmod (nominal modifier)
'agrees'	'in the singular'	'That latter language agrees with Gondi in the singular , but uses the masculine and not the neuter form to denote the plural of nouns which denote women and goddesses .'
'agrees'	'in gender and number'	'In the former case , the participle which forms the tense agrees in gender and number with the object .'
'agrees'	'with Gujarati'	'The oblique form agrees with Gujarati .'
'agrees'	'as in Gondi'	'In Bhili we find forms such as innen bala , thy son , where the possessive pronoun agrees with the qualified noun in the same way as in Gondi .'
		Relation: nsubj (nominal subject)
'agree'	'Adjectives'	'Adjectives agree with their nouns in gender and number , but do not alter with the case of the noun .'
'agree'	'Other forms'	'Other forms mainly agree with Konkani .'
'agree'	'The numerals'	'The numerals often agree very closely with those in use in the Kuki-Chin group .'

Table 1: Frame elements of the AGREEMENT frame

of different FEs of the frame and its relation to other frames. For the purpose of identification of FEs, we have relied on a usage based data driven approach. The procedure we have opted for is as follows:

As a first step, a set of lexical units were identified manually for each of the frame candidates (e.g. agree.v is one potential LU for the AGREEMENT frame). Next, all example sentences from the LSI data containing any of the potential lexical units for a given frame were gathered in a text file. These examples were then parsed using the Stanford Dependency Parser [15], and the constituents within parses were then grouped by the relation type e.g. all (head, dependent) constituent pairs for the relation type 'advmod (i.e. adverbial modifier)' were grouped together. From these groupings, the entries where the head of the relation constituent contained any of the lexical units of a particular frame were separated and saved in a text file. Such a grouping basically gives us all string segments which were used at a particular relational position for a given relation in the whole corpus for each semantic frame. Next, the string constituents were manually observed to determine if an FE is required to capture the information contained within a particular string segment.

Lets take an example to better explain the above given procedure. For the AGREEMENT frame, all sentences containing the lemma 'agree' (considering that agree.v, agreement.n are two lexical units) were collected and parsed as explained above. Table 1 below lists a few selected entries of (Head, Dependent, Sentence) for various relation types.

In addition to general understanding of the frame, the usage based examples given in Table were used to design the following FEs for the AGREEMENT frame.

- Participant_1: Based on the 'nsub' relation and intended to record the first participant of the agreement.
- 'Participant_2': Based on the relation nmod (the argument followed by the keyword 'with') and intended to record the second participant of the agreement.
- 'Grammatical_Category': Based on the relation nmod (the argument usually followed by the keyword 'in') and intended to record the grammatical category on which the two participants agree.
- 'Degree': Based on the relation 'advmod' (e.g. the modifiers 'strongly', 'loosely', etc) and intended to record the degree of agreement.
- 'Frequency': Based on the relation 'advmod' (e.g. the modifiers 'often', 'sometimes', etc.) and intended to record the frequency of agreement.
- 'Language_Variety': General understanding, intended to record the language/variety for which the agreement is being talked about.
- 'Reference_Language': Based on the relation nmod (the argument usually followed by the keyword 'as'), and intended to record the reference language if any.
- 'Condition': Based on the relation 'advcl', and intended to record the condition of agreement if any.

The same procedure was used to identify FEs for all the developed frames.

4.4 Current Status of LingFN

Table 2 shows the current status of LingFN in figures. As can be seen we have developed around 100 frames in total with 32 FEs, around 360 LUs, and more than 2,800 annotated example sentences.

Frame type	# frames	Used-by links	Peripheral FEs	LUs	Annotated examples
Event frames	5	171	16	25	1 858
Filler frames	94	154	16	335	948
Total	99	325	32	360	2806

Table 2: LingFN statistics

5 Applications of LingFN

On the application side, among others, we intend to use LingFN in two ongoing linguistic research projects for automatic extraction of the information encoded in descriptive grammars in order to build extensive typological databases to be used in large-scale comparative linguistic investigations.

The area of automatic linguistic information extraction is very young, and very little work has been previously reported in this direction. Virk et al. [20] report on experiments with pattern and syntactic parsing based methods for automatic linguistic

information extraction. Such methods seem quite restricted and cannot be extended beyond certain limits. We believe a methodology based on the well-established theory of frame semantics is a better option as it offers more flexibility and has proved useful in the area of information extraction in general. The plan is to develop a set of frames for the linguistic domain, annotate a set of descriptive grammars with BFN frames extended by the newly built frame set, train a parser using the annotated data as a training set, and then use the parser to annotate and extract information from the other, unannotated descriptive grammars.

However, in the present paper we limit ourselves to a description of the first part (i.e., development of new frames), and we leave the other tasks (annotations of grammars, training of a parser, and information extraction) as future work.

6 Conclusions and Future Work

Using a combination of the frame-by-frame and lemma-by-lemma framenet development approaches, we have reported on the development of LingFN, an English framenet for the linguistic domain. Taking it as a three stage process, we have described the methodologies for the frame identification, FE identification, and frame development processes. The frames were developed using a frame editor, and the resulting frames have been stored as lexical semantic entries accessible through a general-purpose computational lexical infrastructure.

This is ongoing work, and in the future we plan initially to build additional frames to cover all of the LSI corpus, and subsequently to extend the LingFN beyond this data in order to build a wider-scale resource covering basically the whole linguistic domain.

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 2018-2010 by the Joint Programming Initiative in Cultural Heritage and Global Change, Digital Heritage and Riksantikvarieämbetet, Sweden.

References

1. Baker, C.F., Fillmore, C.J., Lowe, J.B.: The Berkeley FrameNet project. In: Proceedings of ACL/COLING 1998. pp. 86–90. ACL, Montreal (1998), http://dx.doi.org/10.3115/980845.980860
2. Borin, L., Dannélls, D., Forsberg, M., Kokkinakis, D., Toporowska Gronostaj, M.: The past meets the present in Swedish FrameNet++. In: 14th EURALEX International Congress. pp. 269–281. EURALEX, Leeuwarden (2010)

3. Borin, L., Dannélls, D., Forsberg, M., Kokkinakis, D., Toporowska Gronostaj, M.: The past meets the present in Swedish FrameNet++. In: 14th EURALEX International Congress. pp. 269–281. EURALEX, Leeuwarden (2010)
4. Borin, L., Forsberg, M., Lyngfelt, B.: Close encounters of the fifth kind: Some linguistic and computational aspects of the Swedish FrameNet++ project. Veredas 17(1), 28–43 (2013)
5. Borin, L., Forsberg, M., Olsson, L.J., Uppström, J.: The open lexical infrastructure of Språkbanken. In: Proceedings of LREC 2012. pp. 3598–3602. ELRA, Istanbul (2012)
6. Borin, L., Toporowska Gronostaj, M., Kokkinakis, D.: Medical frames as target and tool. In: FRAME 2007: Building Frame Semantics Resources for Scandinavian and Baltic Languages. (Nodalida 2007 Workshop Proceedings). pp. 11–18. NEALT, Tartu (2007)
7. Chiarcos, C.: Ontologies of linguistic annotation: Survey and perspectives. In: Proceedings of LREC 2012. pp. 303–310. ELRA, Istanbul (2012)
8. Fillmore, C.J.: Frame semantics and the nature of language. Annals of the New York Academy of Sciences 280(1), 20–32 (1976), http://dx.doi.org/10.1111/j.1749-6632.1976.tb25467.x
9. Fillmore, C.J.: Scenes-and-frames semantics. No. 59 in Fundamental Studies in Computer Science, North Holland Publishing, Amsterdam (1977)
10. Fillmore, C.J.: Frame semantics. In: Linguistic Society of Korea (ed.) Linguistics in the Morning Calm, pp. 111–137. Hanshin Publishing Co., Seoul (1982)
11. Gildea, D., Jurafsky, D.: Automatic labeling of semantic roles. Computational Linguistics 28(3), 245–288 (2002), http://dx.doi.org/10.1162/089120102760275983
12. Grierson, G.A.: A Linguistic Survey of India, vol. I–XI. Government of India, Central Publication Branch, Calcutta (1903–1927)
13. Hasegawa, Y., Lee-Goldman, R., Kong, A., Akita, K.: Framenet as a resource for paraphrase research. Constructions and Frames 3(1), 104–127 (2011)
14. Malm, P., Ahlberg, M., Rosén, D.: Uneek: A web tool for comparative analysis of annotated texts. In: Proceedings of the IFNW 2018 Workshop on Multilingual FrameNets and Constructicons at LREC 2018. pp. 33–36. ELRA, Miyazaki (2018)
15. Manning, C.D., Surdeanu, M., Bauer, J., Finkel, J., Bethard, S.J., McClosky, D.: The Stanford CoreNLP natural language processing toolkit. In: Proceedings of ACL 2014. pp. 55–60. ACL, Baltimore (2014), http://www.aclweb.org/anthology/P/P14/P14-5010
16. Ponzetto, S.P., Strube, M.: Exploiting semantic role labeling, wordnet and wikipedia for coreference resolution. In: Proceedings of HLT 2006. pp. 192–199. ACL, New York (2006), http://www.aclweb.org/anthology/N/N06/N06-1025
17. Shen, D., Lapata, M.: Using semantic roles to improve question answering. In: Proceedings of EMNLP-CoNLL 2007. pp. 12–21. ACL, Prague (2007), http://www.aclweb.org/anthology/D/D07/D07-1002
18. Surdeanu, M., Harabagiu, S., Williams, J., Aarseth, P.: Using predicate-argument structures for information extraction. In: Proceedings of ACL 2003. pp. 8–15. ACL, Sapporo (2003), http://www.aclweb.org/anthology/P03-1002
19. Torrent, T.T., Salomão, M.M.M., Matos, E.E.d.S., Gamonal, M.A., Gonçalves, J., de Souza, B.P., Gomes, D.S., Peron-Corrêa, S.R.: Multilingual lexicographic annotation for domain-specific electronic dictionaries: The Copa 2014 FrameNet Brasil project. Constructions and Frames 6(1), 73–91 (2014)
20. Virk, S., Borin, L., Saxena, A., Hammarström, H.: Automatic extraction of typological linguistic features from descriptive grammars. In: Proceedings of TSD 2017. Springer (2017)
21. Wu, D., Fung, P.: Semantic roles for SMT: A hybrid two-pass model. In: Proceedings of HLT-NAACL 2009. pp. 13–16. ACL, Boulder (2009), http://dl.acm.org/citation.cfm?id=1620853.1620858