LingFN: Towards a Framenet for the Linguistics Domain

Per Malm[1], Shafqat Mumtaz Virk[2], Lars Borin[2], Anju Saxena[3]
[1]Department of Scandinavian Languages, Uppsala University, Sweden
[2]Språkbanken, University of Gothenburg, Sweden
[3]Department of Linguistics and Philology, Uppsala University, Sweden
per.malm@nordiska.uu.se, shafqat.virk@svenska.gu.se, lars.borin@svenska.gu.se, anju.saxena@lingfil.uu.se

Abstract

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 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, e.g., medicine, soccer, and tourism. In this paper, we report on our experiments and first results on building a framenet to cover the terms and concepts encountered in descriptive linguistic grammars. A contextual statistics based approach is used to judge the polysemous nature of domain-specific terms, and to design new domain-specific frames. The work is part of a more extensive research undertaking where we are developing NLP methodologies for automatic extraction of linguistic information from traditional linguistic descriptions to build typological databases, which otherwise are populated using a labor intensive manual process.

Keywords: domain-specific framenet, information extraction, frame semantic parsing, lexical resource, South Asian linguistics

1. Introduction

Frame semantics is a theory of meaning in language introduced by Charles Filmore and his colleagues (Fillmore, 1976; Fillmore, 1977; Fillmore, 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.

The development of a corresponding lexico-semantic resource – FrameNet (Baker et al., 1998) – 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. The frames are also linked to each other based on a set of conceptual relations making them a network of connected frames, hence the name FrameNet. BFN has proved to be very useful for automatic shallow semantic parsing (Gildea and Jurafsky, 2002), which has applications in a number of natural language processing (NLP) tasks such as information extraction (Surdeanu et al., 2003), question answering (Shen and Lapata, 2007), coreference resolution (Ponzetto and Strube, 2006), paraphrase extraction (Hasegawa et al., 2011), and machine translation (Wu and Fung, 2009; Liu and Gildea, 2010).

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.[1]

In the context of deploying it in NLP applications, BFN and other framenets have often been criticized for their limited coverage. A proposed reasonable-effort solution to this problem this 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).

In this paper, we report our attempts and initial results of building a domain-specific framenet to cover the concepts and terms used in traditional descriptive linguistic grammars. The descriptive grammars are written by linguists in the course of investigating, describing and recording various linguistic characteristics of the target language at the phonological, morphological, syntactic, and semantic levels. For this purpose, linguistics has developed a rich set of specific terms and concepts (e.g. *inflection*, *agreement*, *affixation*, etc.) Useful collections of such terms are provided,

[1]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 (Borin et al., 2010; Borin et al., 2013).

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

e.g., by *GOLD*,[3] the *SIL glossary of linguistic terms*,[4] the *CLARIN concept registry*,[5] and *OLiA* (Chiarcos, 2012).

A minority of these terms are used only in linguistics (e.g., *tense* n.), and in many cases, non-linguistic usages are 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 cannot be expected to cover their usage in linguistics, which we will see below is indeed the case. This means we need to build new frames, identify their triggers and frame elements, and find examples in order to cover them and make them part of the general framenet if we are to extend the coverage. This exactly is one of the major objectives of the experiments we report in this paper.

The work we report on here is part of a more extensive endeavor, where attempts are being made to build methodologies for automatic extraction of the information encoded in descriptive grammars and to build typological databases. The area of automatic linguistic information extraction is very young, and very little work has been previously reported in this direction. Virk et al. (2017) report on experiments with pattern and semantic 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 linguistics-specific frames, annotate a set of descriptive grammars with BFN frames extended by the newly built frame set, train a parser using the annotated data as training set, and then use the parser to annotate and extract information from the other, unannotated descriptive grammars. However, in this paper we limit ourselves to 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.

The rest of the paper is structured as follows: In Section 2, we briefly describe the data that we are using, while Section 3 contains methodological description. Section 4 outlines the frames that we have developed so far and their structure, while the conclusions and an outline of future work follow in Section 5.

[3]`http://linguistics-ontology.org/`

[4]`http://glossary.sil.org`

[5]`https://www.clarin.eu/ccr`

2. The Data

The Linguistic Survey of India (LSI) (Grierson, 1903–1927) 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. Under the supervision of George A. Grierson, the survey resulted into a detailed report comprising 19 volumes of around 9500 pages in total. The survey covered 723 linguistic varieties representing 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). For each major variety it provides (1) a grammatical sketch (including a description of the sound system); (2) a core word list; and (3) text specimens (including a glossed translation of the *Parable of the Prodigal Son*). 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,[6] it is the most comprehensive resource available on South Asian languages, and since it is the major data source in our bigger project, it is natural for us to use it as a starting point for the development of the linguistic framenet, but in the future we plan to extend our range and use other publically available digital descriptive grammars.

3. Methodology

In this section, we describe our methodology at two levels: (1) framenet development; (2) frame development. At the framenet level, there are at least four different types of methodologies which have been discussed in literature. These are the (1) Lexicographic Frame-by-Frame; (2) Corpus-Driven Lemma-by-Lemma; (3) Full-Text; and (4) Domain-by-Domain strategies. In our case, the corpus-driven approach (2) is best suited to our purposes, as our project objectives demand us to cover the available corpus first, and then extend our resource to the domain in general. So we opt to use this approach and build new frames as and when necessary while working with the corpus.

The corpus is in our case the text data of the LSI, i.e., grammar sketches – excluding tabular data (e.g., inflection tables) and text specimens – which have been imported and made searchable using *Korp*, a versatile open-source corpus infrastructure (Borin et al., 2012; Hammarstedt et

[6]The language data for the LSI were collected around the turn of the 20th century, hence obviously reflecting the state of these languages of more than a century ago. However, we know that many grammatical characteristics of a language are quite resistant to change (Nichols, 2003), much more so than vocabulary. In order to get an understanding of the usefulness of the LSI for our purposes, we sampled information from a few of the sketches in order to see how well the LSI data reflect modern language usage. Our results show that while some of the lexical items listed in the LSI are not used today in everyday speech, most other information is still valid for the modern language.

al., 2017a; Hammarstedt et al., 2017b).[7] Currently, the LSI "corpus" comprises about 1.3 MW, and contains data about around 550 linguistic varieties that we identified during the pre-processing step.

At the frame development level, we need to decide when and what domain-specific frames we need to design. Since we are using a domain specific corpus-driven approach, a general rule could be to develop new frames for domain-specific terms describing domain-specific events, concepts, objects and relations etc. But then the question is how to decide which terms are domain-specific and which frames are triggered by them. 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, and for deciding when we need to design a new domain-specific frame, we propose a methodology in the next section and then turn to an illustration of this methodology with an example in the following section.

3.1. Semiautomatic Uniqueness Differentiation

Semiautomatic Uniqueness Differentiation (SUDi) is an approach which can be used to judge the polysemous nature of a given lemma based on the unique contextual attributes of the lemma (Malm et al., forthcoming). This involves five steps: (i) collect sentences containing the polysemous forms from a corpus; (ii) sort these according to usage (general or linguistics-domain specific) into two text files; (iii) annotate the files using a parser/tagger of your choice, preferably one that produces XML which is needed next; (iv) run the XML files through the software *Uneek*; and (v) interpret the result.

With the LingFN project still in the starting blocks, we are also considering other approaches to polysemy disambiguation, both quantitative and qualitative, e.g. Drouin (2003) and Ruppenhofer et al. (2016). These are not discussed here for practical reasons.

Uneek is a web based linguistic tool that may be used to perform an automatic distributional analysis on polysemous forms, on which result it applies set operations, e.g. $A \bigcap B$. It takes two XML files as input. Next, it performs the uniqueness differentiation, i.e. it lists the difference between the files (in set notation $A - B$ and $B - A$). Uneek provides two kinds of statistics: (i) the raw frequencies for each linguistic unit specified in the XML for the A file and for the B file (POS, dependencies, etc.); and (ii) the unique linguistic units for the A file and for the B file.

If Uneek fails to find unique forms for one of the files, then there is no formal support for polysemy. But if it does, one needs to interpret the result.

The uniqueness of a linguistic unit in one domain does not necessarily lead to its infelicity in the other; this must be validated by a linguist. The interpretation is based on proof by contradiction using grammaticality judgements. First you take the linguistic unit that is unique in the context of the polysemous item in one of the files, and place it in the context of the polysemous item in the other file. If this switch results in a reading that is deemed illicit in the tested domain (here marked with #), then you get *positive* formal support to your intuition that the polysemous form may be split into different frames. If the linguistic unit works fine in the other context, then you get negative formal support for polysemy. Paraphrasing Firth (1957): you shall know the difference between two polysemous words by the company one of them constantly rejects.

Step (v) is methodologically problematic since linguists do not always agree on what use should be deemed illicit or not. We do not pretend to have a solution to this difficulty. However, an assessment based on a unique distributional difference is somewhat better than one without any at all.

For our purposes, we are using SUDi to differentiate between two senses: (1) Linguistics Domain Sense (*Ling*); (2) General Domain Sense (*Gen*). For now, we are considering two types of data for the uniqueness differentiation. Either we compare *Ling* forms with all the cases of *Gen* forms found in LSI, or we sort out *Ling* forms and test them against the examples for the LUs in BFN. The last suggestion may seem strange at first since the descriptive statistics would be way off. Yet, since the example sentences of the LUs in BFN exhibit the full range of combinatorial variation (Ruppenhofer et al., 2016, 21), we may use this smaller set in order to find unique clues to domain specific differences. This latter choice is exemplified in the next section.

3.2. An Example

Here we present a methodological example case to illustrate how we motivate a domain specific frame in case of polysemy. We use SUDi to test the assumption of polysemy between *Gen* domain PLACING verbs and *Ling* domain PLACING verbs. We analyze the lemmas based on POS, the surface form words, and dependencies in given order.

A corpus query for *insert*, *place*, and *put*, which are the base form of the verbal lexical units of the BFN PLACING frame, yielded 1 475 hits.[8] 530 of these were assessed to belong to the *Ling* domain.

Moving on to the uniqueness differentiation of POS, we get results indicative of polysemy. The unique POS for the BFN sentences are shown in Table 1, where no unique POS exists for the *Ling* domain PLACING verbs.

Based on the observations in Table 1, we may test how well these unique units work in the *Ling* domain. Let us begin with testing the possessive pronouns in the BFN Example 1 against Example 2 in the *Ling* domain.

(1) *Eadmer*$_i$ inserted them at this point into *his*$_i$ Historia Novorum. (BFN)

Yet, the following invented example indicates that neuter possessive pronouns are not ill suited for the *Ling* domain:

[7] http://spraakbanken.gu.se/swe/forskning/infrastruktur/korp/distribution and https://github.com/spraakbanken/korp-frontend/

[8] There were also one occurrence of *heap* and three of *lay*, but these are excluded for practical reasons.

GENERAL DOMAIN *insert*		GENERAL DOMAIN *place*		GENERAL DOMAIN *put*	
PRP$ 'Possessive pronoun'	10	PRP$ 'Possessive pronoun'	13	JJR 'adjective comparative'	2
WRB 'Wh-adverb'	3	MD 'Modal'	7	WP 'Wh- pronoun'	2
–	–	JJR 'adjective, comparative'	3	JJS 'adjective superlative'	1
–	–	JJS 'adjective, superlative'	1	–	–

Table 1: Some unique features for Gen domain PLACING verbs

Gen *insert*		Gen *place*		Gen *put*	
into	9	place	18	his	15
his	6	on	14	she	15
he	5	he	7	her	11
through	5	them	7	against	10
under	5	has	6	he	7
text	4	under	6	through	7
's	3	against	5	my	6
computer	3	from	5	's	5
left	3	her	5	arm	5
new	3	should	5	said	5

Table 2: Top ten unique PLACING words in the Gen domain

(2) $\textit{Some verb}_i$: a noun is put after $\left\{\begin{array}{l}\#\ \textit{her}_i\\ \#\ \textit{his}_i\\ \textit{its}_i\end{array}\right\}$ base.

This is to be expected since grammatical units are inanimate, thus lacking real agency. A reasonable explanation for why animate possessive pronouns do not occur in the *Ling* domain could be a consistent lack of AGENTS, but for this we need additional proof from the analysis of dependencies.

Next, we observe in Table 1 that superlative and comparative adjectives are unique for the *Gen* domain. A comparison between invented Examples 3a and b below, reveals that these forms seem strange modifiers to *Ling* PLACING words as opposed to *Gen* PLACING words.[9]

(3) a. Goats are put $\left\{\begin{array}{l}\text{closest to}\\ \text{closer to}\\ \text{close to}\end{array}\right\}$ the barn. (GEN)

b. Subjects are put $\left\{\begin{array}{l}\#\ \text{closest to}\\ \#\ \text{closer to}\\ \#\ \text{close to}\end{array}\right\}$ verbs. (LING)

We suspect that anyone consulting a grammar for the placement of the subject in a declarative clause would be rather disappointed to find the inexact answer in Example 3.

[9]However, it is not hard to come up with instances outside our corpus, as also noted by an anonymous reviewer. For instance, it is sometimes observed about certain classes of adjectives that they occur closer to their head noun than some other classes. Similarly, complex affixal morphologies are often described in terms of position classes, where the positions are defined in relation to the stem morph. Again, the use of *closer* and *closest* will come natural in this case.

Moving on to the uniqueness differentiation of words in Table 2, we find that the *Ling* PLACING LUs seem to have restrictions on what may fill the role of GOAL. A linguistic unit may be placed, put, or inserted *before*, *after*, *between*, *at the end of* or *in the beginning of* another linguistic unit. But what about other instantiations of GOALS?

In the *Ling* domain PLACING FEs are not put *into*, *through*, *under*, *on*, or *against* another FE. Notice also in Table 2 the personal pronouns, the present tense contraction *'s*, the modal *should*, and the auxiliary *has*. These observations coupled with the unique distribution of modals presented in Table 1 provide clues for the additional tests. For instance, the linguistic descriptions in LSI do not contain certain modals or non-present tense forms. Arguably, this depends on the factual general claims of the rule-like descriptions. Using modals or complex tense forms while stating a grammatical rule would most likely render the reader confused. See invented Examples 4a–b below.

(4) a. Nouns $\left\{\begin{array}{l}\#\ \text{will}\\ \#\ \text{would}\\ \#\ \text{might}\\ \text{can}\\ \text{may}\end{array}\right\}$ be put after verbs.

b. Nouns $\left\{\begin{array}{l}\#\ \text{are being put}\\ \#\ \text{had been put}\\ \#\ \text{have been put}\\ \#\ \text{were put}\\ \text{are put}\end{array}\right\}$ after verbs.

Last, we look at the uniqueness differentiation of dependencies. The result indicate polysemy and some of the unique distributions are presented in Table 3.

The fact that *Ling place* uniquely contains copulas and that the sentences from the *Ling put* domain uniquely contains 165 passive nominal subjects indicate one particular thing: a lack of active voice in the *Ling* domain. This fact taken together with the temporal restrictions noted in Example 4 and the lack of personal possessives in Table 1 motivates a manual assessment of the Ling domain sentences. The assessment confirms three things of the Ling domain in LSI: (i) verbs are mostly expressed in the passive voice, (ii) the clause is always in the indicative mood, and (iii) always lacks an expressed AGENT, e.g. *by the speaker*. If the voice is active, it is a case of anthropomorphism where a linguistic unit is given agency, e.g. *causal verbs inserts an a after the verb*. There are 35 such cases, all found with *insert*.

In summary, by using SUDi, we have found formal support for a domain specific Linguistic PLACING frame. This is

GENERAL DOMAIN PLACING LUs		LINGUISTIC DOMAIN PLACING LUs	
NMOD:POSS ‘possessive nominal modifier’ (LU=*place*)	17	NSUBJPASS ‘passive nominal subject’ (LU=*put*)	165
NMOD:NPMOD ‘NP as adverbial modifier’ (LU=*insert*)	3	NEG ‘negation’ (LU=*insert*)	12
NMOD:TMOD ‘temporal modifier’ (LU=*put*)	1	COP ‘copula’ (LU=*place*)	4
–		DE:PREDET ‘predeterminer’ (LU=*insert*)	1

Table 3: Some unique dependencies for PLACING LUs in the Gen and Ling domain

strengthened by the interpretation of the results presented in table 1–3 provided by Uneek.

4. Developed Linguistics Domain Frames

Using the methodology described in the previous section, we have developed a few frames specific to the linguistic domain listed in the appendix together with frame triggers, frame elements, and example sentences from our LSI corpus. The following table provides some statistics about the newly developed frames:

Types	Number of types
Frames	12
Core and non-core frame elements	74
Annotated example sentences	156
Lexical units	106

5. Conclusions and Future Work

We have proposed a methodology to judge the polysemous nature of lemmas in a given corpus, and to find their domain-specific occurrence. The decision to build a new domain-specific frame is based on the observation and analysis of the contextual terms that co-occur with a candidate lemma. Using this methodology we have motivated and developed a set of linguistic domain specific frames, and in the future we would like to extend this set. Once we have enough frames, we will start to annotate descriptive grammars with these frames, and then train a parser using the annotated grammars as training data. The parser is then to be used to annotate more grammars and extract linguistic feature values from the annotated texts.

Like all corpus-based methods, Uneek and the results coming out of it are completely dependent on the representativeness of the corpus used. Nevertheless, using it has provided some useful clues to linguistics domain specific word usages, which have formed the basis for our first attempts to devise domain specific frames for the text found in descriptive grammars, as presented in the appendix.

6. Acknowledgements

The work presented here has been financially supported by the Swedish Research Council through its funding of the projects *South Asia as a linguistic area? Exploring big-data methods in areal and genetic linguistics* (2015–2019; contract 421-2014-969) and *Swe-Clarin* (2014–2018; contract 821-2013-2003), the Swedish Foundation for International Cooperation in Research and Higher Education (STINT) through its Swedish-Brazilian research collaboration program (2014–2019; contract BR2014-5860), and the University of Gothenburg, its Faculty of Arts and its Department of Swedish, through their truly long-term support of the Språkbanken research infrastructure.

7. Bibliographical References

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The Berkeley FrameNet project. In *Proceedings of ACL/COLING 1998*, pages 86–90, Montreal. ACL.

Borin, L., Toporowska Gronostaj, M., and Kokkinakis, D. (2007). Medical frames as target and tool. In *FRAME 2007: Building Frame Semantics resources for Scandinavian and Baltic languages. (Nodalida 2007 workshop proceedings)*, pages 11–18, Tartu. NEALT.

Borin, L., Dannélls, D., Forsberg, M., Kokkinakis, D., and Toporowska Gronostaj, M. (2010). The past meets the present in Swedish FrameNet++. In *14th EURALEX International Congress*, pages 269–281, Leeuwarden. EURALEX.

Borin, L., Forsberg, M., and Roxendal, J. (2012). Korp – the corpus infrastructure of Språkbanken. In *Proceedings of LREC 2012*, pages 474–478, Istanbul. ELRA.

Borin, L., Forsberg, M., and Lyngfelt, B. (2013). Close encounters of the fifth kind: Some linguistic and computational aspects of the Swedish FrameNet++ project. *Veredas*, 17(1):28–43.

Chiarcos, C. (2012). Ontologies of linguistic annotation: Survey and perspectives. In *Proceedings of LREC 2012*, pages 303–310, Istanbul. ELRA.

Drouin, P. (2003). Term extraction using non-technical corpora as a point of leverage. *Terminology*, 9(1):99–115.

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

Fillmore, C. J., (1977). *Scenes-and-frames semantics*. Number 59 in Fundamental Studies in Computer Science. North Holland Publishing, Amsterdam.

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

Firth, J. R. (1957). A synopsis of linguistic theory 1930–55. In *Studies in Linguistic Analysis (special volume of the Philological Society)*, pages 1–32. The Philological Society, Oxford.

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

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

Hammarstedt, M., Borin, L., Forsberg, M., Roxendal, J., Schumacher, A., and Öhrman, M. (2017a). Korp 6 – Användarmanual [Korp 6 – User manual]. Technical Report GU-ISS 2017-02, University of Gothenburg, Gothenburg. `http://hdl.handle.net/2077/53096`.

Hammarstedt, M., Roxendal, J., Öhrman, M., Borin, L., Forsberg, M., and Schumacher, A. (2017b). Korp 6 – Technical report. Technical Report GU-ISS 2017-01, University of Gothenburg, Gothenburg. `http://hdl.handle.net/2077/53095`.

Hasegawa, Y., Lee-Goldman, R., Kong, A., and Akita, K. (2011). Framenet as a resource for paraphrase research. *Constructions and Frames*, 3(1):104–127.

Liu, D. and Gildea, D. (2010). Semantic role features for machine translation. In *Proceedings of COLING 2010*, pages 716–724, Beijing. ACL.

Malm, P., Ahlberg, M., and Rosén, D. (forthcoming). 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*, Miyazaki. ELRA.

Nichols, J. (2003). Diversity and stability in language. In Brian D. Joseph et al., editors, *The handbook of historical linguistics*, pages 283–310. Blackwell, Oxford.

Ponzetto, S. P. and Strube, M. (2006). Exploiting semantic role labeling, wordnet and wikipedia for coreference resolution. In *Proceedings of HLT 2006*, pages 192–199, New York. ACL.

Ruppenhofer, J., Ellsworth, M., Petruck, M. R. L., Johnson, C. R., Baker, C. F., and Scheffczyk, J. (2016). *FrameNet II: Extended theory and practice*. ICSI, Berkeley.

Shen, D. and Lapata, M. (2007). Using semantic roles to improve question answering. In *Proceedings of EMNLP-CoNLL 2007*, pages 12–21, Prague. ACL.

Surdeanu, M., Harabagiu, S., Williams, J., and Aarseth, P. (2003). Using predicate-argument structures for information extraction. In *Proceedings of ACL 2003*, pages 8–15, Sapporo. ACL.

Torrent, T. T., Salomão, M. M. M., Matos, E. E. d. S., Gamonal, M. A., Gonçalves, J., de Souza, B. P., Gomes, D. S., and Peron-Corrêa, S. R. (2014). Multilingual lexicographic annotation for domain-specific electronic dictionaries: The Copa 2014 FrameNet Brasil project. *Constructions and Frames*, 6(1):73–91.

Virk, S., Borin, L., Saxena, A., and Hammarström, H. (2017). Automatic extraction of typological linguistic features from descriptive grammars. In *Proceedings of TSD 2017*. Springer.

Wu, D. and Fung, P. (2009). Semantic roles for SMT: A hybrid two-pass model. In *Proceedings of HLT-NAACL 2009*, pages 13–16, Boulder. ACL.

Appendix: Linguistics Domain Frames

Frame	Triggers	Frame elements	Annotated example
AFFIXATION	affix.v, prefixed.a, suffixed.a, affixed.a, infixed.a	**Core:** Morpheme, Morpheme_group, Affix **Non-core:** Degree, Manner, Agent, Condition, Means	[Sometimes]$_{\text{Degree}}$ [it]$_{\text{Morpheme}}$ is [suffixed]$_{\text{LU}}$ to [the genitive]$_{\text{Morpheme}}$
CONJUGATION	conjugate.v, agree.v, inflected.a, change.v, marked.a, conjugated.a, take.v	**Core:** Verb, Grammatical_category, Argument, DNI, Morpheme, Null_morpheme **Non-core:** Degree, Manner, Condition, Means, Agent	[Verbs]$_{\text{Verb}}$ are [regularly]$_{\text{Manner}}$ [inflected]$_{\text{LU}}$ in [person and number]$_{\text{Grammatical\_category}}$.
DECLENSION	put, form	**Core:** Non-verb-word, Grammatical_category, Morpheme, Null_morpheme, DNI **Non-core:** Degree, Manner, Agent, Purpose, Condition	[Adjectives]$_{\text{Non-verb-word}}$ are not [inflected]$_{\text{LU}}$.
DERIVATION	derived.a, changed.a, transform.v, take.v	**Core:** Word, Derivational_morpheme, Null_morpheme, Part_of_speech, DNI, Condition **Non-core:** Degree, Means	[It]$_{\text{Word}}$ [must]$_{\text{Degree}}$ be [derived]$_{\text{LU}}$ from [a verb substantive with a negative prefix]$_{\text{Derivational\_morpheme}}$
GRAMMATICAL_CASE	nominative.n, accusative.n, dative.n, ablative.n, genitive.n, vocative.n, locative.n, instrumental.n, oblique.n, agent.n	**Core:** Grammatical_case **Non-core:** Descriptor	The [accusative$_{\text{Grammatical_case}}$ is the case of the object .
INFLECTION	inflected.a, conjugate.v, agree.v, decline.v, marked.a, conjugated.a, change.v, take.v, put.a	**Core:** Word, Word_group, Inflectional_morpheme, Grammatical_category, CNI **Non-core:** Degree, Manner, Condition, Purpose, Means	[Verbs]$_{\text{Word\_group}}$ are [regularly]$_{\text{Manner}}$ [inflected]$_{\text{LU}}$ [in person and number]$_{\text{Grammatical\_category}}$
MORPHOLOGICAL_ENTITY	suffix, affix, prefix, infix	**Core:** Morphological_entity **Non-core:** Descriptor, Type, Constituent_parts	Siki is the [corresponding]$_{\text{Descriptor}}$ [suffix]$_{\text{Morphological\_entity}}$ [of the object]$_{\text{Constituent\_parts}}$.
SYNTACTIC_CONFIGURATION	put.a, put.v, arrange.v, stand.v, placed.a, inserted.a, follow.v, precede.v, come.v	**Core:** Syntactic_position, Syntactic_unit_1, Syntactic_unit_2 **Non-core:** Degree, Manner, Condition	[The verb]$_{\text{Syntactic\_unit\_1}}$ [usually]$_{\text{Degree}}$ [comes]$_{\text{LU}}$ [last in the sentence]$_{\text{Syntactic\_position}}$.
SYNTACTIC_ROLE	subject.n, object.n, predicate.n, adjunct.n, clause.n	**Core:** Syntactic_role **Non-core:** Descriptor, Type, Constituent_parts	The usual order of words is [subject]$_{\text{Syntactic\_role}}$, [object]$_{\text{Syntactic\_role}}$, verb.
VERB INDEXING	agree.v, inflected.a, change.v, marked.a, take.v	**Core:** Verb, Grammatical_category, Argument **Non-core:** Condition, Degree, Means, Manner	[The verb]$_{\text{Verb}}$ [agrees]$_{\text{LU}}$ [in gender and person]$_{\text{Grammatical\_category}}$ [with the object]$_{\text{Argument}}$, [when the object is in the form of the nominative]$_{\text{Condition}}$.
LINGUISTIC_ENTITY	suffix.n, affix.n, prefix.n, infix.n, conjunction.n, cardinal.n, determiner.n, preposition.n, adjective.n, adverb.n, verb.n, modal.n, noun.n, predeterminer.n, particle.n, infinitive.n, interjection.n, gerund.n, participle.n, ordinal.n, nominative.n, ablative.n, accusative.n, dative.n, genitive.n, vocative.n, locative.n, instrumental.n, oblique.n, agent.n	**Core:** Linguistic_entity **Non-core:** Descriptor, Type, Constituent_parts	This is an example of the[dative]$_{\text{Linguistic\_entity}}$ [of possession]$_{\text{Descriptor}}$