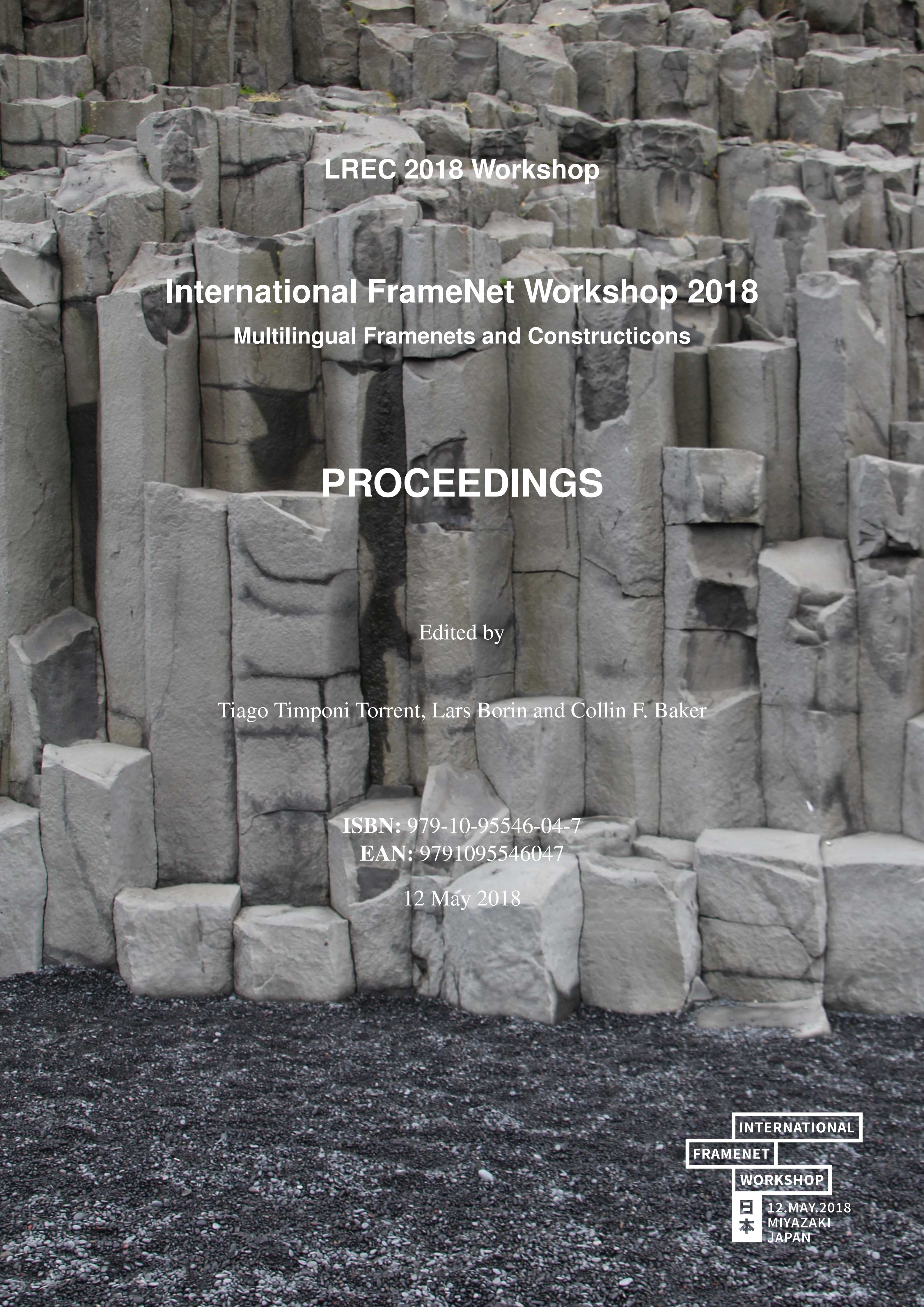


LREC 2018 Workshop

International FrameNet Workshop 2018

Multilingual Framenets and Constructicons

PROCEEDINGS

Edited by

Tiago Timponi Torrent, Lars Borin and Collin F. Baker

ISBN: 979-10-95546-04-7
EAN: 9791095546047

12 May 2018

INTERNATIONAL
FRAMENET
WORKSHOP
日本
12.MAY.2018
MIYAZAKI
JAPAN

Proceedings of the LREC 2018 Workshop
International FrameNet Workshop 2018: Multilingual Framenets and Constructicons

12 May 2018 – Miyazaki, Japan

Edited by Tiago Timponi Torrent, Lars Borin and Collin F. Baker

Front cover photo: *Reynisfjara rocks!* by Lars Borin

Acknowledgments: The arrangement of this workshop has been made possible by a grant for Brazilian-Swedish research collaboration awarded jointly by CAPES (Brazil) and STINT (Sweden) to the FrameNet/Constructicon teams at the Federal University of Juiz de Fora and the University of Gothenburg.

Organizing Committee

- Collin F. Baker (FrameNet Project, International Computer Science Institute, Berkeley, California)
- Lars Borin (Språkbanken, Department of Swedish, University of Gothenburg, Sweden)
- Benjamin Lyngfelt (Department of Swedish, University of Gothenburg, Sweden)
- Tiago Timponi Torrent (Federal University of Juiz de Fora, Brazil)

Program Committee

- Collin F. Baker, International Computer Science Institute
- Hans Boas, University of Texas, Austin
- Dana Dannells, University of Gothenburg
- Lars Borin, University of Gothenburg
- Gerard de Melo, Rutgers University
- Ellen Dodge, International Computer Science Institute
- Jerome Feldman, International Computer Science Institute
- Markus Forsberg, University of Gothenburg
- Karin Friberg Heppin, Independent researcher
- Normunds Gruzitis, University of Latvia
- Richard Johansson, University of Gothenburg
- Dimitrios Kokkinakis, University of Gothenburg
- Russell Lee-Goldman, Google Inc.
- Ely Matos, Federal University of Juiz de Fora
- Srini Narayanan, Google Inc.
- Pierre Nugues, Lund University
- Martha Palmer, University of Colorado
- Miriam R L Petruck, International Computer Science Institute
- Josef Ruppenhofer, Institute for the German Language, Mannheim
- Nathan Schneider, Georgetown University
- Francis Tyers, Higher School of Economics, Moscow
- Tiago Timponi Torrent, Federal University of Juiz de Fora
- Shafqat Mumtaz Virk, University of Gothenburg

Preface

The *International FrameNet Workshop 2018* brought together researchers in Frame Semantics and Construction Grammar, two areas which have traditionally been interrelated, but which have been developing somewhat independently in recent years. It is also addressed at language technology researchers working with language resources based on Frame Semantics or Construction Grammar. The workshop follows on from similar joint meetings in Berkeley, California in 2013 (IFNW 2013, sponsored by the Swedish FrameNet group) and in Juiz de Fora, Brazil in 2016 (IFNW 2016, sponsored by FrameNet Brasil), and will cover the rapidly unfolding developments in both areas and recent research on their interconnections.

Charles J. Fillmore and Paul Kay and their students and colleagues developed the theories of Frame Semantics and Construction Grammar in parallel over a period of several decades. Both have been of interest to many linguists, psychologists, computer scientists, and others, with most people tending to be more interested in one than the other. This workshop will attempt to bring together researchers working on Construction Grammar with those working on Frame Semantics, both from a theoretical (linguistic) and more practical (language technology) perspective, highlighting the interconnections of the two theories, their relation to other theories of semantics and syntax, as well as their deployment in concrete natural language processing applications. This workshop will also provide a forum for reporting on cross-lingual and multilingual research on Frame Semantics and Construction Grammar around the world.

In the call for papers we invited submissions discussing theoretical questions related to Frame Semantics and Construction Grammar (CxnG), especially in a multilingual context, and preferably empirically based (on corpus studies and/or natural language processing applications), such as the following:

- What counts as a construction? What counts as a frame?
- Are the schemas of CxnG necessarily different from FrameNet frames? If so, how and why? Are Frames/Schemas an adequate semantic representation for CxnG? What constructions are implicit in ordinary FrameNet-style annotation? Are relations between constructions basically the same as relations between frames?
- To what extent are semantic frames language universals? How should cross-linguistic differences in frames be represented and studied?
- To what extent are constructions the same across languages? How can we make useful cross-linguistic comparisons between semantically similar constructions such as correlatives, conditionals, causatives, etc.?
- How can research devoted mainly to either Frame Semantics or Construction Grammar contribute to the growth of both approaches?

We also welcomed (2) reports on language resources based on Frame Semantics (framenets) or Construction Grammar (constructicons) being developed and made freely available in any language, including reports on annotation using the new Multilingual FrameNet annotation tool described below.

In addition, we especially invited (3) reports on applications of Frame Semantics and Construction Grammar, including both Frame Semantic parsers/semantic role labeling systems and Construction Grammar parsers and end-to-end systems.

Submissions for the workshop followed the LREC extended abstract format of 3–4 pages of text, plus additional pages of references, as needed. However, and differently from the main LREC conference, all submissions for the IFNW 2018 workshop had to be anonymous. We are happy to say that the submissions together covered many of the topics listed in the call for papers. Each submission was anonymously reviewed by three members of the program committee. The final workshop program contains 13 presentations, 6 oral presentations and 7 posters, and the present volume contains full-length versions of the workshop papers, extended and revised according to the reviewers' comments.

Tiago Timponi Torrent, Lars Borin and Collin F. Baker March 2018

Workshop program

Opening Session

09.00–09.15 Collin F. Baker: *Welcome and introduction to the workshop*

09.15–10.30 **Oral session 1**

Luca Gilardi and Collin F. Baker
Learning to align across languages: Toward Multilingual FrameNet

Tiago Timponi Torrent, Michael Ellsworth, Collin F. Baker
and Ely Edison da Silva Matos
The Multilingual FrameNet shared annotation task: A preliminary report

Piek Vossen, Antske Fokkens, Isa Maks and Chantal van Son
Towards an Open Dutch FrameNet lexicon and corpus

10.30–11.00 **Coffee break**

11.00–11.45 **Poster session**

Normunds Gruzitis, Gunta Nespore-Berzkalne and Baiba Saulite
Creation of Latvian FrameNet based on Universal Dependencies

Stephen Wright Horn, Alastair Butler, Iku Nagasaki and Kei Yoshimoto
Deriving mappings for FrameNet construction from a parsed corpus of Japanese

Alexandre Costa, Maucha Gamonal, Vanessa Paiva, Natália Marção,
Simone Peron-Corrêa, Vânia Almeida, Ely Matos and Tiago Torrent
*FrameNet-based modeling of the domains of tourism and sports for
the development of a personal travel assistant application*

Kyoko Ohara, Daisuke Kawahara, Satoshi Sekine and Kentaro Inui
*Linking Japanese FrameNet with Kyoto University Case Frames
using crowdsourcing*

Gabriel Marzinotto, Frederic Bechet, Geraldine Damnati and Alexis Nasr
Sources of complexity in semantic frame parsing for information extraction

Sanni Nimb
The Danish FrameNet lexicon: Method and lexical coverage

Shafqat Mumtaz Virk and K.V.S. Prasad
Towards Hindi/Urdu Framenets via the Multilingual FrameNet

11.45–13.00 **Oral session 2**

Per Malm, Shafqat Mumtaz Virk, Lars Borin and Anju Saxena
LingFN: Towards a framenet for the linguistics domain

Waad Alhoshan, Riza Batista-Navarro and Liping Zhao
*A FrameNet-based approach for annotating natural language
descriptions of software requirements*

Per Malm, Malin Ahlberg and Dan Rosén
Uneek: A web tool for linguistic analysis

13.00 **Closing**

Table of contents

A FrameNet-based Approach for Annotating Natural Language Descriptions of Software Requirements

Waad Alhoshan, Riza Batista-Navarro, Liping Zhao
School of Computer Science, University of Manchester, United Kingdom
waad.alhoshan@postgrad.manchester.ac.uk
{riza.batista, liping.zhao}@manchester.ac.uk

Abstract

As most software requirements are written in natural language, they are unstructured and do not adhere to any formalism. Processing them automatically—within the context of software requirements engineering tasks—thus becomes difficult for machines. As a step towards adding structure to requirements documents, we exploited frames in FrameNet and applied them to the semantic annotation of software descriptions. This was carried out through an approach based on automated lexical unit matching, manual validation and harmonisation. As a result, we produced a novel corpus of requirements documents containing software descriptions which have been assigned a total of 242 unique semantic frames overall. Our evaluation of the resulting annotations shows substantial agreement between our two annotators, encouraging us to pursue finer-grained semantic annotation as part of future work.

Keywords: Semantic Frames, FrameNet, Corpus Annotation, Software Requirements, Requirements Engineering

1. Introduction

Software requirements play a pivotal role in all system design phases. Requirements are generally written in natural language, and therefore are unstructured (Ferrari et al., 2017a). This however presents a challenge to Requirements Engineering (RE) tasks, e.g. requirements analysis, which often necessitate the organisation and management of requirements in a systematic manner (Dick et al, 2017). While certain RE tasks (e.g., modelling) could benefit from automated analysis, this can only be facilitated if some structure is applied to the otherwise unstructured natural language requirements contained in software descriptions (Ferrari et al., 2017b).

One way by which we can add structure to software descriptions written in natural language is by attaching machine-readable semantic metadata that captures meaning. In documents from the general and scientific domains, this often corresponds to named entities, e.g., proper names of persons, places, diseases or chemical compounds. Software descriptions however do not allude to such proper names as often and instead mention generic if not abstract concepts (e.g., account creation, file deletion) and the participants involved (e.g., user, system). As shown in early work by Belkhouche and Kozma (1993) and Rolland and Priox (1992), capturing meaning contained in requirements can be approached by using *semantic frames*: coherent structured representations of concepts (Petruck, 1997). These representations are based on the theory of *frame semantics* proposed by Fillmore (1977) whose work formed the basis of FrameNet, an online computational lexicon that catalogues detailed information on semantic frames[1] (Baker et al., 1998). For every frame it contains, FrameNet specifies the following: frame title, definition, frame elements (i.e., participants) and lexical units, i.e., words that evoke the frame. The concept of creation, for example, is encoded in FrameNet as a frame entitled *Creating*, with frame elements pertaining to *Creator*, *Created_entity* and *Beneficiary* (among many others). Importantly, lexical units that signify the concept is also provided, each of which is represented as a combination of their lemmatised form and part-of-speech (POS) tag (e.g., *assemble.v*, *create.v* where *v* stands for verb). Such a frame can then be applied on a piece of text (such as in Example 1) to represent, in a structured manner, the creation idea that is being conveyed. Containing over 1,200 such frames, FrameNet has become an invaluable resource to the NLP research community.

Example 1:
[The system] Creator [generates] Creating lexical unit [records of user activities] Created_entity [each time] Frequency [the user logs into the system] Cause.

Recent studies in RE have explored the application of FrameNet frames to software requirements acquisition and analysis. For example, Jha and Mahmoud (2017) employed semantic frames (automatically extracted by the SEMAFOR semantic role labeller[2]) as features in training machine learning-based models for categorising user reviews of mobile applications. Meanwhile, Kundi and Chitchyan (2017) proposed a technique for gathering requirements that employed FrameNet frames as the basis of linguistic patterns for generating use cases at the early stages of RE. They specifically made use of the *Agriculture* frame to demonstrate their approach.

We consider FrameNet as a rich repository of semantic metadata that can be added to requirements documents in order to add structure to them. In this work, we seek to employ FrameNet as the basis of a scheme for capturing the meaning of software descriptions. To this end, we adopt FrameNet semantic frames in annotating software requirements in a corpus of documents written in natural language. To the best of our knowledge, our work is the first attempt to investigate FrameNet as a means for annotating meaning within requirements documents. In this way, we are enriching them with semantic metadata and hence incorporating structure into them. As a result, we have produced and made publicly available a resource for the perusal of other members of the research

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

[2] http://www.cs.cmu.edu/~ark/SEMAFOR/

community: the FrameNet-annotated FN-REQ[3] corpus of natural language requirements documents.

The rest of this paper is organised as follows. Section 2 describes our methods for collecting software requirements documents and annotating them based on the semantic frames contained in FrameNet. In Section 3, we present and analyse results of our annotation. Lastly, we present our conclusions and plans for future work in Section 4.

2. Methodology

In this section, we present the methods we carried out in order to construct a corpus of documents containing sentences of software requirements, and to subsequently annotate them according to FrameNet.

2.1 Document Selection

Our goal is to gather a document set consisting of different types of software requirements. As a preliminary step, we formed a Google search query containing keywords such as "software description", “natural language requirements" and "software requirements specification". Furthermore, we employed snowball sampling and found additional requirements from various sources such as web blogs, research articles (together with their corresponding datasets), lecture materials and industrial/commercial documents. This step resulted in the collection of 34 requirements documents varying in length. The NLTK tool[4] for sentence boundary detection was then applied on the 34 documents. After manually verifying the results, a total of 1,148 sentences[5] were obtained (corresponding to 21,012 tokens).

2.2 Annotation Procedure

The annotation was carried out in a semi-automatic manner. This was facilitated by the two main steps described as follows.

2.2.1 Evoking Frames by Lexical Unit Matching

With the intention of making the annotation process more efficient, we developed a simple method for automatically matching words in the software descriptions in our corpus against lexical units contained in FrameNet, in order to evoke candidate semantic frames. The tokens contained in the requirements documents were lemmatised and assigned part-of-speech (POS) tags using NLTK. For every description, we attempt to match each token (together with its lemma and POS tag) against lexical units in FrameNet, via the application programming interface (API) available in NLTK[6]. We note that only particular types of FrameNet lexical units were considered by this matching method, namely: all verbs and any expressions pertaining to time (e.g., "beforehand"), condition (e.g., "in case", "otherwise"), additional action (e.g., "further"), inclusion (e.g., "inclusive"), exclusion (e.g., "excluding"), contradiction (e.g., "nevertheless"), causation (e.g., "because of") and purpose (e.g., "in order"). The selection of these types was informed by our observations on the linguistic styles often used in writing software requirements. Through this process, we were able to evoke candidate semantic frames that denote the meaning of the requirements in our documents.

2.2.2 Validation

Deciding which FrameNet semantic frames capture the meaning expressed in software descriptions was performed manually in order to maximise accuracy. For this task, we employed two annotators. The first annotator (Annotator A) is a requirements engineer with five years of experience in the IT industry. The second annotator (Annotator B) is one of the authors of this paper and is a PhD candidate whose study is focussed on the use of NLP techniques to support RE tasks.

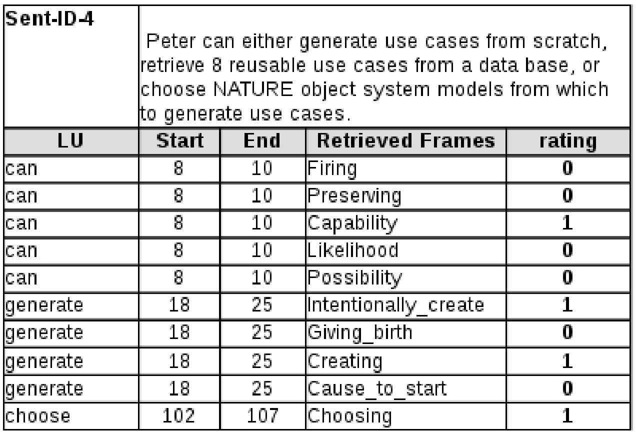
Provided with candidate frames obtained in the previous step, the annotators were asked to confirm whether they capture the meaning of a given software description or not. This validation process was carried out in accordance with the guidelines we developed which drew inspiration from the FrameNet annotation scheme proposed by (Baker, 2017). Over a four-week period, both annotators were trained in applying these guidelines on the annotation of a set of software descriptions from documents other than those in our corpus. Afterwards, the entire corpus of 34 documents—together with the candidate semantic frames retrieved in the previous step—was presented to each of Annotators A and B for annotation. We provide Table 1 to show an example of the details that are presented to an annotator and the kind of judgement that he/she is expected to provide. At the top row of the table is a sample software description. The first column (LU) lists the lexical units matched by the method described in Section 2.2.1. The second and third columns (Start and End) indicate the location of the corresponding lexical unit in terms of character offsets—useful information in cases where a lexical unit appears multiple times within a description. The fourth column (Retrieved Frames) lists the titles of the frames linked with the matched lexical units and are thus considered as candidate frames for annotating the given description. The annotator indicates in the last column his/her judgement on whether a candidate frame applies to the software description (rating = 1) or not (rating = 0). Both annotators completed this task for all 1,148 software descriptions in our corpus.

[3] Read as "fine req"

[4] http://www.nltk.org/api/nltk.html

[5] Identified based on sentence delimiters such as the full stop. Not all of these however are sentences in the strict sense; some are phrases. They all however pertain to software descriptions, thus we use "descriptions" rather than "sentences" in the rest of this paper.

[6] http://www.nltk.org/howto/framenet.html



Sent-ID-4	Peter can either generate use cases from scratch, retrieve 8 reusable use cases from a data base, or choose NATURE object system models from which to generate use cases.			
LU	**Start**	**End**	**Retrieved Frames**	**rating**
can	8	10	Firing	**0**
can	8	10	Preserving	**0**
can	8	10	Capability	**1**
can	8	10	Likelihood	**0**
can	8	10	Possibility	**0**
generate	18	25	Intentionally_create	**1**
generate	18	25	Giving_birth	**0**
generate	18	25	Creating	**1**
generate	18	25	Cause_to_start	**0**
choose	102	107	Choosing	**1**

Table 1. A sample software description from the corpus. An annotator is presented with the automatically matched lexical units, their character offset locations and the titles of the frames linked with them. He/she then indicates whether the frames apply to the requirements (rating = 1) or not (rating = 0). (NB: The second instance of "generate" is also presented to the annotator but excluded here for brevity.)

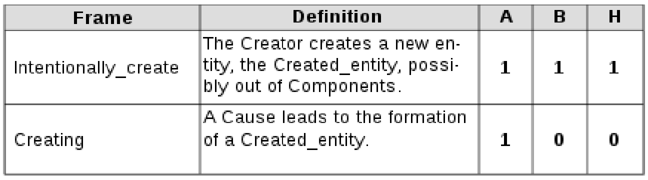
3. Results and Discussion

In this section, we discuss the results of the methodology described above by providing details on inter-annotator agreement and reasons behind annotator discrepancies. We then describe additional steps that were taken in order to prepare the corpus for publication. After presenting attributes of the resulting corpus in terms of annotation frequencies, we discuss a few suggestions on how our proposed annotation method can be useful to members of the research community within the context of RE tasks.

3.1 Inter-annotator Agreement

In order to assess the consistency of annotations between our two annotators, we evaluated inter-annotator agreement based on Cohen's kappa coefficient (McHugh, 2012) as well as the harmonic mean of recall and precision, i.e., F-score. We obtained "substantial" agreement[7] according to Cohen's kappa (72.81%). Furthermore, after determining the number of true positives, false positives and false negatives (by treating the annotations from Annotator B as gold standard and those from Annotator A as response) and micro-averaging over all the documents in our corpus, we obtained an F-score of 80.89%. These results indicate that there is a more than satisfactory level of consistency between our two annotators, implying that their annotations can be considered as highly reliable.

Nevertheless, we investigated the reasons of discrepancy between our two annotators. We found that these are mostly due to close semantic relationships between certain semantic frames. FrameNet, for example, contains a *Creating* and an *Intentionally_create* frame, both of which would be retrieved by our automated lexical unit matching method—and thus presented to an annotator—for a description containing the word "generate" as a verb. As these two frames have similar lexical units and are linked by hyponymy (where *Intentionally_create* has *Creating* as its parent frame), Annotator A could select one frame while Annotator B might select the other (or both, as shown in the example in Table 2). Aiming to produce annotations that are of the highest quality as possible, we resolved these discrepancies, as described in the next section, prior to publishing the annotated corpus.



Frame	Definition	A	B	H
Intentionally_create	The Creator creates a new entity, the Created_entity, possibly out of Components.	1	1	1
Creating	A Cause leads to the formation of a Created_entity.	1	0	0

Table 2. A case where Annotator A's judgements on which frames apply to the the word "generate" (in the software description in Table 1), are in disagreement with those of Annotator B. This can be attributed to the hyponymic relationship between the *Intentionally_create* and *Creating* frames. The last column is for recording the results of harmonisation (H).

3.2 Preparation of the Final Corpus

In order to produce the final set of annotations, we harmonised the judgements provided by our two annotators, addressing the primary cause of discrepancies discussed in the previous section. From the set of semantic frames for which the annotators were in disagreement, the following instances were revisited by Annotator B: (1) where the FrameNet frame that she selected as being most relevant to a description is semantically related to the one selected by Annotator A; and (2) where multiple—presumably semantically related—frames were selected for a word in a description. Annotator B reviewed information pertinent to the frames in question, e.g., the definitions and descriptions provided in FrameNet, examples of annotations in the FrameNet corpus[8], as well as the judgements provided by Annotator A. In cases where she is convinced that Annotator A's judgements were more correct, she modified her own annotations; otherwise, she kept her original judgements. She also ensured that only one frame is assigned to a given word (i.e., the matched lexical unit), choosing the one that best captures the meaning of a description (as she understands it), while also reviewing the definitions and examples that are available in FrameNet. The outcome of this process formed the basis of the final set of annotations in our corpus.

3.3 Frequency Analysis

After harmonisation of manually provided judgements, we performed frequency analysis over the final set of annotations, the results of which are presented in Table 3. Alongside these we also provide the frequency of annotations resulting from our automated lexical unit matching method, as the reader might be interested in seeing how much improvement was obtained after manual validation and harmonisation. As one can expect, the automated method for matching lexical units introduced a considerable amount of noise. Firstly, the matching of

[7] As stipulated in Landis and Koch (1977)

[8] Refer to Language Resource Reference

tokens (with their lemmatised forms and POS tags) against FrameNet lexical units does not have perfect accuracy as the POS tagger that we utilised was assigning the wrong POS tag to tokens in a few cases. Secondly, for a given word from a description, e.g., "generate", our method would have retrieved all frames that are associated with the "generate" lexical unit regardless of the sense (e.g., *Intentionally_create*, *Giving_birth*, *Creating*, *Cause_to_start*). This would have resulted in a significant number of false positives, i.e., frames that are irrelevant to a given software description. These issues were however rectified during manual validation and subsequently, during harmonisation.

In our final set of annotations, only frames with rating = 1 (after manual validation and harmonisation) were included. We can observe from Table 3 that out of the 408 semantic frames retrieved through automated lexical unit matching, 166 (40.7%) were eliminated during manual validation and harmonisation, and thus were not included in the final set. There was also a significant drop in terms of the average number of frames assigned to each software description (from 8.82 per description to only 2.21).

	Automated lexical unit matching	**Final set of annotations**
Total number of unique frames	408	242
Total number of unique lexical units	372	340
Average number of frames per software description	8.82	2.21

Table 3. Frequency analysis over the final set of annotations in the FN-REQ corpus. For comparison, we also provide the frequency of annotations obtained through automated lexical unit matching (prior to manual validation and harmonisation).

Our corpus can be considered as densely annotated, with semantic frames assigned to 88.4% of the total number of descriptions (1,015 out of 1,148). Annotations were encoded in a standoff manner, i.e., separately from the documents that were annotated. While the requirements documents were stored following an extended version of the schema proposed by (Ferrari et al., 2017), the annotations were encoded according to the FrameNet format (Baker, 2017).

3.4 Potential Applications

The utilisation of frames in FrameNet to attach semantic metadata to software descriptions—as demonstrated in this work—could potentially facilitate the (partial) automation of certain requirements engineering tasks. For instance, similarities between requirements statements written in natural language can be automatically detected or measured on the basis of the semantic frames assigned to each of them. This in turn can enable *traceability*, i.e., establishing relationships or groupings between requirements and effectively, the software systems they pertain to (Zogaan et al., 2017). Additionally, attaching semantic metadata derived from FrameNet to requirements statements makes them machine-readable and hence more searchable. A software engineer developing requirements for a new system can thus find existing requirements of relevance in a more efficient and systematic manner. In this way, the *reusability* of existing requirements can be enhanced, hence avoiding unnecessary duplication of efforts (Alonso-Rorís et al., 2016).

4. Conclusion and Future Work

In this work, we demonstrated how semantic frames can be applied to the annotation of software descriptions. Along the way, we produced FN-REQ corpus, which we have made publicly available, together with other associated resources (e.g., annotation guidelines, the script that automates matching of FrameNet lexical units), at https://data.mendeley.com/datasets/s7gcp54wbv/1 .

As we were progressing with the manual annotation process described in this work, both annotators observed that there are words in some descriptions which to them clearly pertain to software requirements, but however cannot be assigned any of the frames in FrameNet. For example, it is now typical for software requirements to mention the process of logging into a system, often signified by the verb "log" (as in Example 1 in Section 1). However, none of the frames in the most recent version of FrameNet conveys this concept. This is not a surprise as FrameNet is a general vocabulary and was not designed to cater to specific domains. However, for our purposes of supporting requirements engineering tasks as part of downstream applications, it is worth investigating how many of such requirements in our corpus are currently not covered by FrameNet, in order to assess if there is scope for extending it through the proposal of new additional frames. This is part of our ongoing work. Furthermore, we are in the process of extending our FN-REQ corpus with more requirements documents, while we also carry out finer-grained annotation of software descriptions by labelling frame elements as well. In our future work, we shall exploit the corpus in the context of RE tasks, specifically in detecting traceability and reusability of software requirements.

Acknowledgement

We thank Mohammed Homaid for the time and effort he spent on the annotation of our corpus. We are also grateful for our reviewers whose comments and feedback helped improve the paper. Waad Alhoshan's PhD is sponsored by the Al-Imam Muhammad Ibn Saud University..

References

Alonso-Rorís, V. M., Álvarez-Sabucedo, L., Santos-Gago, J. M., & Ramos-Merino, M. (2016). Towards a cost-effective and reusable traceability system. A semantic approach. *Computers in Industry*, (83):1-11).

Baker, C. F., Fillmore, C. J., & Lowe, J. B. (1998, August). The berkeley framenet project. In *Proceedings of the 17th international conference on Computational linguistics-Volume 1* (pp. 86-90). Association for Computational Linguistics.

Baker, C. F. (2017). FrameNet: Frame Semantic Annotation in Practice. In *Handbook of Linguistic Annotation* (pp. 771-811). Springer, Dordrecht.

Belkhouche, B., & Kozma, J. (1993). Semantic case analysis of informal requirements. In *Proceedings of the 4th Workshop on the Next Generation of CASE Tools* (pp. 163-181).

Hull, E., Jackson, K., & Dick, J. (2010). Management Aspects of Requirements Engineering. In *Requirements Engineering* (pp. 159-180). Springer London.

Ferrari, A., DellOrletta, F., Esuli, A., Gervasi, V., & Gnesi, S. (2017a). Natural Language Requirements Processing: A 4D Vision. *IEEE Software*, *34*(6), (pp. 28-35), IEEE.

Ferrari, A., Spagnolo, G. O., & Gnesi, S. (2017). PURE: A Dataset of Public Requirements Documents. In *Requirements Engineering Conference (RE), 2017 IEEE 25th International* (pp. 502-505). IEEE.

Fillmore, C. J. (1977). Scenes-and-frames semantics. *Linguistic structures processing*, *59*, 55-88.

Jha, N., & Mahmoud, A. (2017). Mining user requirements from application store reviews using frame semantics. In *International Working Conference on Requirements Engineering: Foundation for Software Quality*(pp. 273-287). Springer, Cham.

Kundi, M., & Chitchyan, R. (2017). Use Case Elicitation with FrameNet Frames. In *2017 IEEE 25th International Requirements Engineering Conference Workshops (REW)* (pp. 224-231). IEEE.

McHugh, M. L. (2012). Interrater reliability: the kappa statistic. *Biochemia Medica: Casopis Hrvatskoga Drustva Medicinskih Biokemicara / HDMB*, (pp. 276-282).

Petruck, M. R. L. (1997). Frame semantics. Handbook of Pragmatics (Vol. 12, pp. 1-13). Amsterdam: John Benjamins Publishing Company.

Rolland, C., & Proix, C. (1992). A natural language approach for Requirements Engineering. In Advanced Information Systems Engineering (pp. 257-277). Springer, Berlin, Heidelberg.

Landis, J. R., & Koch, G. G. (1977). An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, (pp. 363-374).

Zogaan, W., Sharma, P., Mirahkorli, M., & Arnaoudova, V. (2017). Datasets from Fifteen Years of Automated Requirements Traceability Research: Current State, Characteristics, and Quality. In *Proceedings of the 25th International Requirements Engineering Conference* (pp. 110-121). IEEE.

Language Resource Reference

FrameNet. (2017). The FrameNet project, distributed via International Computer Science Institute in Berkeley, 1.7, URL: https://goo.gl/Nbuqvd

FrameNet-Based Modeling of the Domains of Tourism and Sports for the Development of a Personal Travel Assistant Application

Alexandre Diniz da Costa, Maucha Andrade Gamonal, Vanessa Maria Ramos Lopes Paiva,
Natália Duarte Marção, Simone Rodrigues Peron-Corrêa, Vânia Gomes de Almeida,
Ely Edison da Silva Matos, Tiago Timponi Torrent
Federal University of Juiz de Fora – FrameNet Brasil
Rua José Lourenço Kelmer, s/nº – Campus Universitário – 36036-900 – Juiz de Fora/MG – Brazil
{alexandre.costa, ely.matos, tiago.torrent}@ufjf.edu.br, {duarte.natalia, vania.almeida2017}@letras.ufjf.br,
{mauchaandrade, speronjf, vanessaletrasufjf}@gmail.com

Abstract

This paper presents an enriched frame-based multilingual lexicon covering the domains of Tourism and Sports, which supports a personal travel assistant application – m.knob – developed to help tourists get recommendations of attractions and activities, as well as to communicate with other tourists and service providers, in the context of major international sports events, such as the Summer Olympics. Recommendations are provided through frame-based automated categorization of tourist attractions based on semantic information extracted from tourists' comments on online platforms, which are then matched to semantic information extracted from the input the user inserts in a conversational user interface.

Keywords: Tourism and Sports Modeling, Algorithmic Categorization, m.knob, FrameNet Brasil.

1. Introduction

Events such as the Summer Olympics provide the meeting of people from different parts of the world, who have different interests related to tourist attractions and sports, as well as speak different languages. Therefore, major international events like this one call for multilingual tools that can assist tourists in their choices related to places to eat or visit, sports events to attend, and so on.

Also, planning a trip or leisure activity requires different types of information about a tourist attraction or event. Many travel guides can assist in bringing information about places, how to get there, what to do, or even the temperature and weather conditions at any given time of the year. Likewise, these tools often focus on prominent attractions or more general information that aid in the basic planning for a trip. However, travel guides do not provide specific information that many tourists may need when planning a trip, such as which attraction is better for a rainy day or which museum is interesting for children. This information is either subjective and subject to change or is scattered around the text. While this kind of information may be available on online platforms in the form of comments and reviews posted by users, reading them all is a task incompatible with the dynamism of a trip.

Considering this context, an automatic analysis of these comments could generate more useful information to the tourist, especially if they are made available in an interactive and dynamic platform. It is not only a matter of extracting if the general impression about a certain attraction is positive or negative, an already classic task in Natural Language Processing (NLP), but also to go beyond such classification, bringing more specific information that helps the user make decisions. In addition, this specific information can also help the tourist to choose the sports disciplines, considering the context of the Olympic games, and to find the places where the competitions take place, since sports are also a type of leisure activity searched by tourists in this context.

This work is developed under the m.knob (Multilingual Knowledge Base) project of the FrameNet Brasil Computational Linguistics Laboratory at the Federal University of Juiz de Fora. Such a project is developing a personal travel assistant in the form of a chatbot with which tourists can interact using natural language to get recommendations for attractions, places to eat and leisure activities.

In this context, this paper aims (a) to show how the modeling was carried out, and (b) to present an automated categorization methodology for tourist attractions based on semantic information extracted from comments posted to online platforms. Such a methodology provides for the existence of an analyzer that extracts the semantic information from the comments and translates it into a cluster of frames. The system also generates clusters from the user's inputs and later maps the similarities between the clusters, suggesting attractions and tourist activities that can adhere to the user's interests.

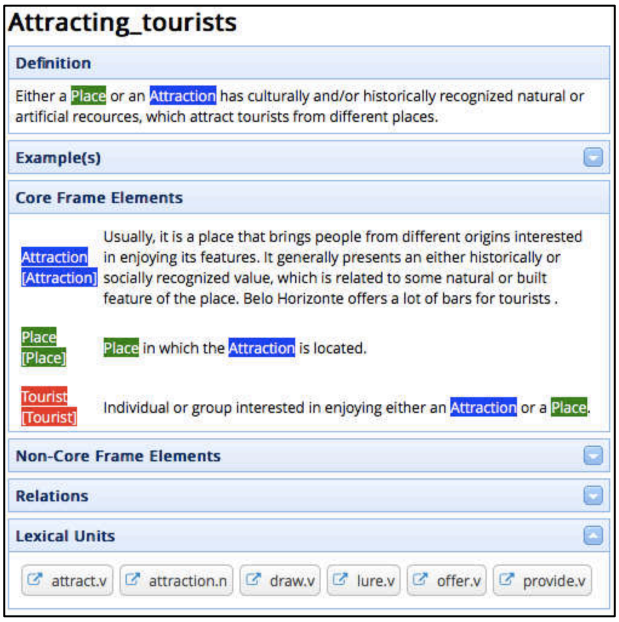
2. Frame Semantics and FrameNets

Frame Semantics is an approach to lexical semantics whose main assumption is that meanings are relativized to scenes (Fillmore, 1977), that is, to frames. Fillmore (1985) proposes an approach to semantics based on language understanding, analyzing the linguistic choices made to produce utterances so that they convey beliefs about the world, experiences, and the way speakers see things. Frames are defined as a system of concepts related in such a way that "to understand one of them, it is necessary to understand the whole structure in which it fits" (Fillmore 1982, p. 111).

The main application of Frame Semantics is FrameNet, a project started in the International Computer Science

Institute (ICSI), by Charles Fillmore, with the purpose of providing, through the exposition of Lexical Units (LUs), the frames evoked by these LUs, identified by the Frame Elements (FEs) that constitute them. By FEs, we mean any semantic role specifically defined in the frame. FEs provide additional information to the semantic structure of the sentence. LUs, in turn, are pairings of lemmas and the frames they evoke (Fillmore, 1982). The analyzes performed on the LUs, therefore, provide us with a description of their syntactic valence properties (grammatical functions and syntagmatic types that co-occur in the syntactic locality of the lexical item) and semantics (frame elements instantiated by these valents).

Figure 1 shows the Attracting_tourists frame, its FEs and LUs. There's also a definition of the frame, as well as one for its core FEs – ATTRACTION, PLACE and TOURIST – and their definitions as well.



Attracting_tourists

Definition

Either a Place or an Attraction has culturally and/or historically recognized natural or artificial recources, which attract tourists from different places.

Example(s)

Core Frame Elements

Attraction [Attraction] Usually, it is a place that brings people from different origins interested in enjoying its features. It generally presents an either historically or socially recognized value, which is related to some natural or built feature of the place. Belo Horizonte offers a lot of bars for tourists .

Place [Place] Place in which the Attraction is located.

Tourist [Tourist] Individual or group interested in enjoying either an Attraction or a Place.

Non-Core Frame Elements

Relations

Lexical Units

attract.v attraction.n draw.v lure.v offer.v provide.v

Figure 1: The Attracting_tourists frame.

LUs evoking this frame include *attract.v, draw.v, lure.v, offer.v and provide.v.* Sentences containing these LUs are annotated in a multiple layer fashion (Frame Element, Grammatical Function and Phrase Type), and show clear examples of basic combinatorial possibilities (valence patterns) for each target LU. Note that, although some of these lemmas may also take part in LUs evoking different frames – such as Cause_motion, Manipulate_into_doing, Offering and Supply, respectively – their sense in the context of sentences (1-5), extracted from travel guides in the FrameNet Brasil corpus, takes the Attracting_tourists frame as a background, not the other frames mentioned above, as indicated by the color code matching the linguistic material in each sentence to the FEs shown in Figure 1.

(1) The mighty Songhua River, running through Harbin from west to east, inevitably ATTRACTS tourists.

(2) Manzanillo initially DREW the interest of international visitors for its excellent fishing.

(3) Alicante Swaying palms and luminous skies, along with some of Spain's best restaurants and tapas bars, LURE visitors to the provincial capital of Alicante.

(4) Few countries OFFER so much to visitors as Brazil.

(5) Kuta, and its progressively upscale neighbors to the north ' Legian , Seminyak , and Kerobokan (as well as Tuban, to the south) PROVIDE an enormous selection of hotels, restaurants, pubs, and shopping choices . INI

Based on FrameNet, lexical resources are being developed for different languages such as German (Boas et al., 2006), Japanese (Ohara et al., 2004), Spanish (Subirats & Petruck, 2003), Chinese (You & Liu, 2005), Swedish (Borin et al., 2010) and Brazilian Portuguese (Salomão, 2009). Similarly to Berkeley FrameNet, FrameNet Brasil follows the same methodology with a team of linguists and computer scientists who are involved in various fields of research, from the construction of lexical resources to the development of applications for natural language understanding. We now turn to one of such applications developed by FrameNet Brasil: m.knob.

3. Multilingual Knowledge Base

Multilingual Knowledge Base (m.knob) is a travel assistant app that offers personalized information to tourists about the specific domains of Tourism and Sports. The alpha version of the app was released during the Rio 2016 Summer Olympics and has been redesigned to include other functions in its beta version.

The app covers three languages – Brazilian Portuguese, English and Spanish – and has two main functions, (i) a chatbot providing recommendations on tourist attractions and activities; and (ii) a semantically enhanced sentence translator algorithm based on frames and qualia relations (Pustejovsky, 1995).

The Tourism domain was modeled in a previous application: the 2014 World Cup Dictionary. Torrent et al. (2014) developed a frame-based trilingual electronic dictionary for the 2014 World Cup, covering the domains of Football, Tourism and the World Cup in the same three languages. The modeling carried out for the Tourism domain (Gamonal, 2013; Gomes, 2014; Souza, 2014) included, at first, 40 frames. For m.knob, it has been revised and improved to cover other aspects of the travel experience, and currently features 58 frames, 16 of which already existed in the Berkeley FrameNet Data Release 1.7. As for the Sports Domain, Costa & Torrent (2017) created 29 new frames and used 4 frames from Berkeley FrameNet 1.7. Currently, the m.knob lexicon comprises a total of 5,152 LUs: 1,671 for Brazilian Portuguese, 2,551 for English, 930 for Spanish.

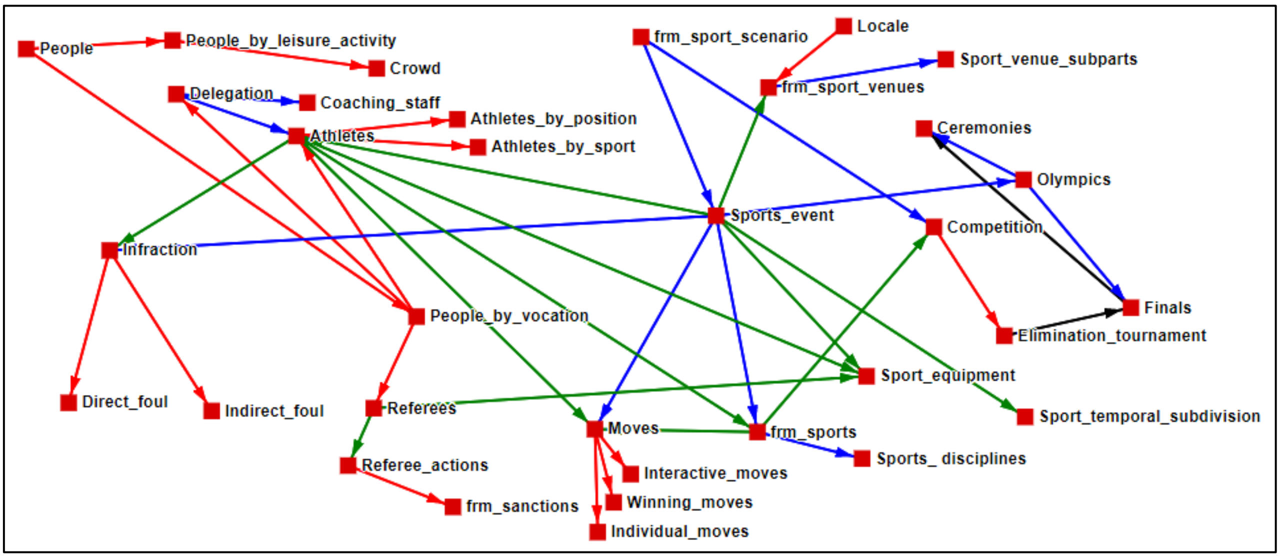


Figure 2: Frames and relations in the Sports domain. Arrow colors indicate the types of relations: Inheritance (red), Using (green), Subframe (blue) and Precedes (black).

The process of modeling the Tourism and Sports domains, besides creating new frames, also led to the enrichment of FrameNet Brasil to the extent that it incorporated new relations to the database. This process is discussed next.

3.1 Modeling the Tourism and Sports Domains

The process of creating and modeling the frames for Tourism and Sports adopted a bottom-up approach and started with the compilation of trilingual corpora related to the domains. Texts were extracted from travel guides and blogs, governmental portals on tourism and on the Olympics, as well as from sports manuals and websites of associations of each Olympic sport. The corpus compilation tool used was Sketch Engine (Kilgarriff et al., 2014).

Next, candidate terms in the corpus were extracted using TermoStat (Drouin, 2003) and the context in which they occur is analyzed to both (i) validate the term as evoking a frame related to the relevant domains, and (ii) expand the list of candidate terms. Example sentences were then analyzed to provide the basis for the proposition of the frames. Finally, the resulting proto-frames were then refined – based on the literature on tourism (Gamonal, 2013) and on the rules of the Olympic sports –, and related to one another in a network, using the frame-to-frame relations originally defined by Berkeley FrameNet (Ruppenhofer et al., 2016). The resulting model for the Sports domain is presented in Figure 2.

Besides the frames and LUs modeling the specific terminology of Sports and Tourism, the m.knob lexicon also contains domain general frames and LUs relevant to the description of tourist and sports attractions. Such frames and LUs were selected from the Berkeley FrameNet data release 1.7 and expanded into Brazilian Portuguese and Spanish. This selection was based on a pilot study in which a corpus of 3,495 comments written in English about 939 tourist locations in San Francisco was analyzed semi-automatically in a three-step procedure:

- first, candidate LUs were automatically extracted from the corpus, by comparing the word forms in the comments to those associated to LUs – and, therefore, frames – in Berkeley FrameNet;
- second, frames were ranked from the most to the least frequent, regardless of the LU evoking them;
- third, annotators in the FrameNet Brasil team manually checked which frames were actually relevant and which of them were irrelevant to the domains.

Among the examples of relevant frames are Stimulus_focus (evoked by LUs such as *great.a, beautiful.a, interesting.a*), Expensiveness (*expensive.a, cheap.a*), Kinship (*son.n, grandfather.n*), People_by_age (*child.n, senior.a*), Locales_by_use (*museum.n, church.n*), Natural_features (*LUs such as beach.n and valley.n*) and so on. Frames were judged as irrelevant mostly when the word forms triggering their recognition by the system should actually point to another frame, or to no frame at all in the context of the comments. The parade examples are the Performers_and_roles (evoked by *be.v*) and the Sex (evoked by have.v) frames. Both *be.v* and *have.v* are very frequent in the comments, but not in the senses of playing some character or having sex, respectively.

The pilot study resulted in the incorporation of 250 Berkeley FrameNet frames to the m.knob lexicon. English LUs evoking those frames were imported into the database from the data release 1.7. Brazilian Portuguese and Spanish LUs are being created in those frames through the regular expand process used in FrameNet Brasil (Torrent & Ellsworth, 2013).

The conceptual structure represented by the m.knob lexicon is a graph. Nodes in this graph include lemmas, LUs, frames and FEs. The arcs in this graph are the several relations between those nodes, such as the frame-to-frame relations currently used by most – if not all – framenets, but also new ones, which were created by FrameNet Brasil, such as FE-to-frame, qualia and metonymy relations.

Because the m.knob lexicon is meant to be used as the basis for a recommendation system and a sentence translator, new relations were added to the database apart from those originally created by Berkeley FrameNet – illustrated in Figure 2 – either to provide more specific links – connecting LUs instead of frames –, or to account for the definition of the entities participating in an event and for the possible metonymic relations between those entities.

The first set of new relations, those connecting LUs, was adapted from Pustejovsky's (1995) qualia (Costa & Torrent, 2017). So far, three different qualia were implemented in the m.knob database: formal, constitutive and telic. The formal quale is used to indicate that a given LU has the same ontological type of another, more generic LU. It is a *is-a* relation and is used to indicate, for example, that *taphouse.n*, *sports bar.n* and *pothouse.n* are a *bar.n*. The constitutive quale indicates that the referent of a given LU functions as a part or content of the referent of another LU. It indicates for example that *bleachers.n* and *field.n* are parts of a *stadium.n*. Finally, the telic quale is used, in m.knob, to indicate either the inherent purpose of an object or the actions prototypically performed by an agent. It is used to indicate, for example, that the *ace.n* in a soccer team usually scores a *goal.n*, but not an *ace.n*, which is prototypically performed by a *tennis player.n*.

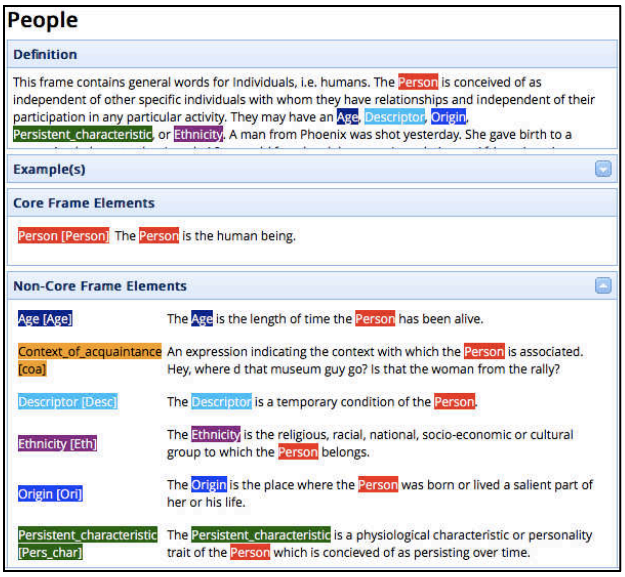
The second set of new relations models the fact that participants in a frame can be defined in terms of other (entity) frames, and also that, in some cases, they can be represented metonymically. Using the Attracting_tourists frame (Figure 1) as an example, an FE-to-frame relation models that the PLACE FE may be defined in terms of the Locale frame, while the TOURIST FE may be defined in terms of the People frame (Figure 3). Additionally, inside the People frame, a FE-to-FE Metonymy relation indicates that the non-core FE ORIGIN, may stand for the core FE PEOPLE (Gamonal, 2017).

Changes as the one just described, allow m.knob to extract, from (6), that the Attracting_tourists frame was evoked in the sentence, because:

- first, an FE-to-frame relation links *city.n*, in the Political_locale frame to the FE PLACE, via the Locale frame;
- second, the Metonymy relation creates a link between *Brazilian.a* and *people.n* – or any other LU in the People frame;
- third, an FE-to-frame relation links *Brazilian.a* to the FE TOURIST, via the People frame.

(6) The city lures Brazilians with beautiful beaches and nice shops.

This kind of structure is then key to m.knob's recommendation system, which will be presented in section 3.2.



People

Definition

This frame contains general words for Individuals, i.e. humans. The Person is conceived of as independent of other specific individuals with whom they have relationships and independent of their participation in any particular activity. They may have an Age, Descriptor, Origin, Persistent_characteristic, or Ethnicity. A man from Phoenix was shot yesterday. She gave birth to a

Example(s)

Core Frame Elements

Person [Person] The Person is the human being.

Non-Core Frame Elements

Age [Age]	The Age is the length of time the Person has been alive.
Context_of_acquaintance [coa]	An expression indicating the context with which the Person is associated. Hey, where d that museum guy go? Is that the woman from the rally?
Descriptor [Desc]	The Descriptor is a temporary condition of the Person.
Ethnicity [Eth]	The Ethnicity is the religious, racial, national, socio-economic or cultural group to which the Person belongs.
Origin [Ori]	The Origin is the place where the Person was born or lived a salient part of her or his life.
Persistent_characteristic [Pers_char]	The Persistent_characteristic is a physiological characteristic or personality trait of the Person which is concieved of as persisting over time.

Figure 3: The People frame

3.2 Automated Categorization of Attractions

Although the collaborative culture of the internet has made subjective assessments of tourist attractions available through diverse tools, this is still not enough for the user to take advantage of this information, given the impossibility of reading all the comments when planning a trip. The application described in this work overcomes these limitations through a categorization algorithm that uses the m.knob lexicon to generate detailed semantic representations of attractions and events.

Based on the algorithmic categorizer, the system parses comments posted to online platforms and extracts the meaning of the candidate words. In a first stage, the set of frames evoked in the comments is gathered. Then, the evoked frames are weighed as to their frequency in the data. In a third step, the frame clusters representing each place are derived and stored in the m.knob database, as well as additional information about the place itself, such as its name, opening hours, location and, very important, its type in the online platform. Such types are stored also in the m.knob lexicon, in the form of LUs such as *bar.n, park.n, beach.n* and so on. Place types are usually the dominant node of formal quale relations, as the ones exemplified in section 3.1.

On the other end, a conversational user interface, namely a chatbot, provides the user with the possibility of entering, in one of the three languages covered by the resource, what she'd like to do. In the final stage, the system provides the tourist with recommendations resulting from a cluster-matching process between the semantic representation generated for the user's input and those generated for the attractions from the analysis of the comments.

As an example, consider that one user enters sentences (7), (8), (9) or (10) to the chatbot system.

(7) Quero passear com a minha família.
I want to go out with my family.
(8) Quero passear com a minha família à noite.
I want to go out with my family tonight.
(9) Quero ter contato com a natureza.
I want to be close to nature.
(10) Quero passear com a minha família em contato com a natureza.
I want to go out with my family to be close to nature.

First, the system extracts the LU candidates from the sentences and finds in the m.knob lexicon the correspondences shown in Table 1.

Br-Pt LU	En Gloss	Frame
passear.v	*go out*	Going_places
família.n	*family*	Kinship
contato.n_1	*be in contact with*	Contacting
contato.n_2	*be close to*	Spatial_contact
natureza.n	*nature*	Natural_features

Table 1: LUs found in sentences (7-10) and the frames they evoke in the m.knob lexicon

Second, using the relations between frames, FEs and LUs described in section 3.1, the system disambiguates the lemmas pointing to more than one LU. In this example, *contato.n 'contact'* is an ambiguous lemma, since it could refer to both an LU in the Contacting frame and one in the Spacial_contact frame. However, in the user input, it appears close to *natureza.n 'nature'*, which evokes the Natural_features frame. Based on that, the system infers that Spatial_contact is more likely, because the distance – in terms of the relations described in 3.1 and also those common to FrameNet, such as Inheritance, Perspective and so on – between this frame and Natural_features is shorter than that between Contacting and Natural_features (see Torrent et al., 2014 for a description of the frame disambiguation system).

Third, the system generates a semantic cluster to represent the user query. In this process, it takes two other kinds of linguistic information into account, besides the LUs found in the query: words that do not evoke frames, but appear both in the user input and in the comments – such as *noite.n 'night'*, for example –, and other LUs evoking the frames in the query – such as *filho.n 'son', pai.n 'father', mother.n 'mãe'*, in the Kinship frame, and *montanha.n 'mountain'* in the Natural_features frame. That way, the system, once again, makes use of the network-like infrastructure of FrameNet to broaden the linguistic bases used for recommendation.

Next, the cluster representing the query is to be matched to those representing places to be recommended. This is made possible by: first, turning the cluster into a graph in which LUs, frames, and other words are nodes and the relations connecting them in the m.knob lexicon are arcs, and, second, by applying spreading activation techniques to this graph to find which of the places in the database is the best fit for the user query (see Matos et al., 2017 for a description of the spreading activation process used in FrameNet Brasil).

For the sake of exemplification, let's assume that the m.knob database has six places which are potentially relevant to queries (7-10). By applying the first three steps described for the analysis of the user query to the comments written about those places – namely, LU candidate extraction, frame disambiguation and semantic cluster generation –, the system derives a semantic cluster representing each place, as shown in Table 2.

Such clusters are also represented as graphs, whose nodes will be activated in the cluster matching process. In the end, the places the system will recommend to the user are those with the highest activation levels achieved based on the user input and how it matches to the semantic representation of the place.

Place_#	LUs	Frames	Other
Place_1	*contato.n_2* *natureza.n*	Spatial_contact Natural_features	
Place_2	*contato.n_2* *natureza.n*	Spatial_contact Natural_features	
Place_3	*passear.v* *família.n*	Going_places Kinship	
Place_4	*passear.v* *família.n*	Going_places Kinship	
Place_5	*passear.v* *família.n*	Going_places Kinship	*noite.n*
Place_6	*contato.n_2* *natureza.n* *passear.v* *família.n*	Spatial_contact Natural_features Going_places Kinship	

Table 2: LUs, frames and other relevant words in the clusters describing Places 1-6 in the m.knob database

Hence, given, for example, the user input in (7), the system would recommend Places 3, 4, 5 and 6, all of them with an activation level of 1.9368, as shown in Figure 4.

Note that the activation process starts by setting the activation value of each LU in the query to 1.000. Then, every time the activation spreads to another node via an arc, this value is reduced. When a node is activated by more than one path, activation values are added up in the final node.

For the user input in sentence (8), once again Places 3, 4, 5 and 6 are activated. However, Place_5 has a higher activation value [1.9611], as shown in Figure 5, and would then be recommended as the best-fit option to the user query. This is so because both the query in (8) and Place_5 feature the word *noite.n*, demonstrating that additional information provided by the user may help the system provide better recommendations.

As for sentence (9), the activation process yields Places 1, 2 and 6 as equally good recommendations. However, if the user input is (10), then all places are activated, but Place_6 gets a higher activation score [3.8710], as shown in Figure 6.

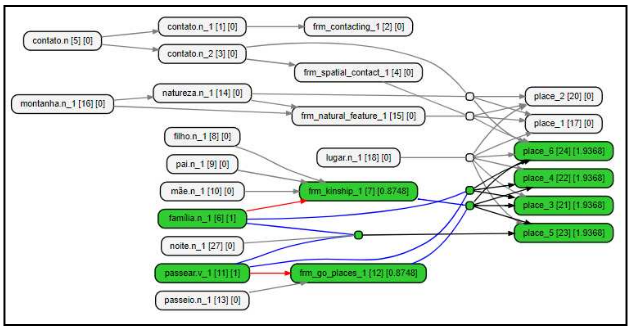


Figure 4: Graph representation of the cluster-matching process between sentence (7) and Places 1 to 6. Nodes in green are activated and numbers in the second pair of square brackets indicate their level of activation.

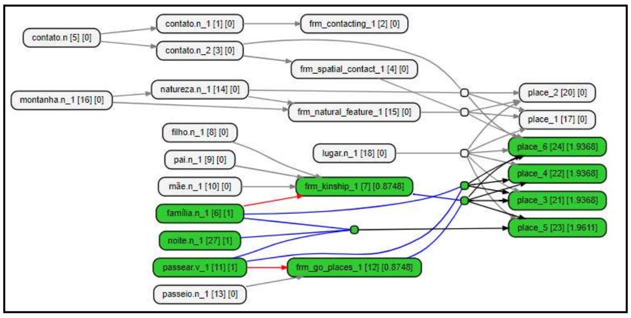


Figure 5: Graph representation of the cluster-matching process between sentence (8) and Places 1 to 6. Nodes in green are activated and numbers in the second pair of square brackets indicate their level of activation.

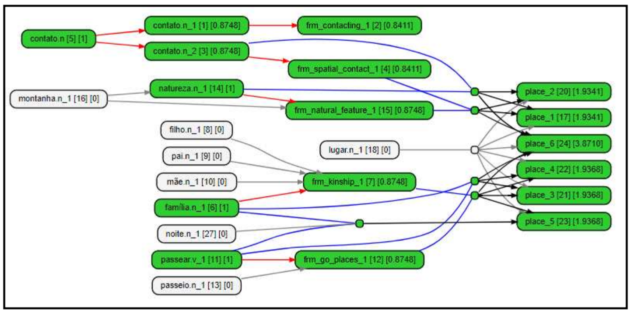


Figure 6: Graph representation of the cluster-matching process between sentence (10) and Places 1 to 6. Nodes in green are activated and numbers in the second pair of square brackets indicate their level of activation.

4. Currently Limitations and Outlook

In this paper, we demonstrated how a domain-specific framenet for Tourism and Sports can be used for providing recommendations for tourists by applying spreading activation techniques to graphs representing the semantics of user inputs to a conversational interface. Currently limitations of the system refer to both the lexicon and the algorithm.

On the side of the lexicon, there's, first, the need to balance the number of LUs for each language. Currently, the number of LUs in Spanish is half of that in Brazilian Portuguese, which, in turn, is 50% lower than that of English LUs. Second, the consistency of the newly created relations in the database must be checked.

On the side of the algorithm, the clusterization process operating on the comments uses n-grams to delimit the scope of the lemma disambiguation process. This is not ideal, since n-grams do not capture the structural relations between the lexical items, and, the m.knob lexicon, on the other hand, models plenty of those relations. In the future, we plan to substitute the use of n-grams by the constructional parser being developed by FrameNet Brasil (Matos et al., 2017).

5. Acknowledgements

The m.knob project is funded by CNPq Grant No. 448990-2014/8 and by FAPEMIG Grant No. CHE-APQ-00471/15.

6. Bibliographical References

Boas, H. C., Ponvert, E., Guajardo, M., and Rao, S. (2006). The current status of German FrameNet. In SALSA workshop at the University of the Saarland (Vol. 14), Saarbrucken, Germany, june.

Borin, L., Dannélls, D., Forsberg, M., Gronostaj, M. T., and Kokkinakis, D. (2010). The past meets the present in Swedish FrameNet++. In *Proceedings of the 14th EURALEX international congress*, pages 269–281, Reykjavik, Icenland, may. European Language Resource Association (ELRA).

Costa, A. D., and Torrent, T. T. (2017). A Modelagem Computacional do Domínio dos Esportes na FrameNet Brasil. In *Proceedings of the 11th Brazilian Symposium in Information and Human Language Technology*, pages 201–208, Uberlândia, Brazil, nov. Sociedade Brasileira de Computação (SBC).

Drouin, P. (2003). Term Extraction Using Non-technical Corpora as a Point of Leverage. *Terminology*, 9(1): 99–115.

Fillmore, C. J. (1977). The case for case reopened. In P. Cole & J. Saddock (eds). *Grammatical Relations*. New York: Academic Press, pp. 59--81.

Fillmore, C. J. (1982). Frame Semantics. In: The Linguistic Society of Korea (org.). *Linguistics in the morning calm*. Seoul: Hanshin, pp. 111--137.

Fillmore, C. J. (1985). Frames and the semantics of understanding. *Quaderni di Semantica*, 6(2):222–254.

Gamonal, M. A. (2013). *COPA 2014 FrameNet Brasil: Diretrizes para a Constituição de um Dicionário Eletrônico Trilíngue a partir da Análise de Frames da Experiência Turística*. M.A. Thesis in Linguistics. Universidade Federal de Juiz de Fora. Juiz de Fora.

Gomes, D. S. (2014). *Frames do Turismo Esportivo no Dicionário Copa 2014_FrameNet Brasil*. M.A. Thesis in Linguistics. Universidade Federal de Juiz de Fora. Juiz de Fora.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovvář, V., Michelfeit, J., Rychlý and P., Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, 1: 7–36.

Matos, E. E., Torrent, T. T., Almeida, V. G., Silva, A. B. L., Lage, L. M., Marção, N. D., and Tavares, T. S. (2017). Constructional Analysis Using Constrained Spreading Activation in a FrameNet-Based Structured Connectionist Model In: *The AAAI 2017 Spring Symposium on Computational Construction Grammar* *and Natural Language Understanding Technical Report SS-17-02*, pages 222–229, Palo Alto, CA, mar. Association for The Advancement of Artificial Intelligence (AAAI).

Ohara, K. H., Fujii, S., Ohori, T., Suzuki, R., Saito, H., and Ishizaki, S. (2004). The Japanese FrameNet Project: An Introduction. In *Proceedings of LREC-04 Satellite Workshop Building Lexical Resources from Semantically Annotated Corpora (LREC 2004)*, pages 9-11, Lisbon, Portugal, may. European Language Resource Association (ELRA).

Pustejovsky, J. (1995). *The Generative Lexicon*. Cambridge, USA, MIT Press.

Salomão, M. M. M. (2009). FrameNet Brasil: um trabalho em progresso. *Calidoscópio*, 7(3):171–182.

Souza, B. C. P. (2014). *Frames de turismo como negócio no Dicionário Copa 2014_FrameNet Brasil*. M.A. Thesis in Linguistics. Universidade Federal de Juiz de Fora. Juiz de Fora.

Subirats, C., and Petruck, M. (2003). Surprise: Spanish FrameNet. In *Proceedings of CIL*, Vol. 17, Prague, Czech Republic, jun.

Torrent, T. T., and Ellsworth, M. (2013). Behind the Labels: Criteria for Defining Analytical Categories in FrameNet Brasil. *Veredas*, 17(1): 44–65.

Torrent, T. T., Salomão, M. M., Campos, F. A., Braga, R. M., Matos, E. E., Gamonal, M. A., Gonçalves, J., Souza, B. C., Gomes, D. S., and Peron-Correa, S. R. (2014). Copa 2014 FrameNet Brasil. In *Proceedings of COLING 2014*, pages 10–14, Dublin, Ireland, aug. Association for Computational Linguistics (ACL).

You, L., and Liu, K. (2005). Building Chinese FrameNet database. In *Proceedings of 2005 IEEE International Conference on Natural Language Processing and Knowledge Engineering*, pages 301–306, Whuan, China, oct-nov. Institute of Electrical and Electronic Engineers (IEEE).

Learning to Align across Languages: Toward Multilingual FrameNet

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

Abstract

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

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

1. The FrameNet Project at ICSI

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

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

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

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

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

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

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

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

2. FrameNet-related Projects for Other Languages

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Table 1: Summary of FrameNet Projects by Language

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

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

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

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

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

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

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

3. Towards an Aligned Multilingual FrameNet

3.1. Overview

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

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

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

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

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

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

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

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

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

3.2. Aligning FrameNets

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

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

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

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

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

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

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

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

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

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

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

4. Conclusions

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

5. Acknowledgments

This material is based in part upon work supported by the U.S. National Science Foundation under Grant No. (1629989). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

6. Bibliographical References

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Computational Linguistics.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Creation of Latvian FrameNet based on Universal Dependencies

Normunds Gruzitis, Gunta Nespore-Berzkalne, Baiba Saulite
University of Latvia, Institute of Mathematics and Computer Science, Raina blvd. 29, Riga, Latvia
{normunds.gruzitis,gunta.nespore,baiba.valkovska}@lumii.lv

Abstract

This paper presents a work in progress, creating a FrameNet-annotated text corpus for Latvian. This is a part of a larger project which aims at the creation of a multilayered corpus, anchored in cross-lingual state-of-the-art syntactic and semantic representations: Universal Dependencies (UD), FrameNet and PropBank, as well as Abstract Meaning Representation (AMR). For annotating the FrameNet layer, we use the latest frame inventory of Berkeley FrameNet, while the annotation itself is done on top of the underlying UD layer. Thus, the annotation of frames and frame elements is guided by the dependency structure of a sentence, instead of the phrase structure. We strictly follow a corpus-driven approach, meaning that lexical units in Latvian FrameNet are created only based on the annotated corpus examples. Since we are aiming at a medium-sized still general-purpose corpus for a less-resourced language, an important aspect that we take into account is the variety and balance of the corpus in terms of genres, domains and lexical units.

Keywords: FrameNet, Universal Dependencies, Latvian

1. Introduction

Natural language understanding (NLU) systems rely, explicitly or implicitly, on syntactic and semantic parsing of text. State-of-the-art parsers, in turn, typically rely on supervised machine learning which requires substantial language resources – syntactically and semantically annotated text corpora, and extensive linked lexicons.

In the industry-oriented research project "Full Stack of Language Resources for Natural Language Understanding and Generation in Latvian" (Gruzitis et al., 2018), we are creating a balanced text corpus with multilayered annotations, adopting widely acknowledged and cross-lingually applicable representations: Universal Dependencies (UD) (Nivre et al., 2016), FrameNet (Fillmore et al., 2003), PropBank (Palmer et al., 2005) and Abstract Meaning Representation (AMR) (Banarescu et al., 2013).

The UD representation is automatically derived from a more elaborated manually annotated hybrid dependency-constituency representation (Pretkalnina et al., 2016). This also ensures that paragraphs, sentences and tokens are correctly and uniformly split and represented in the standard CoNLL-U data format (see Table 1) before the FrameNet annotation begins. All the annotation layers are afterwards merged, based on the document, paragraph, sentence and token identifiers. The FrameNet annotations are manually added, guided by the underlying UD annotations (see Figure 1). Consequently, frame elements are represented by the root nodes of the respective subtrees instead of text spans; the spans can be easily calculated from the subtrees. The PropBank layer is automatically derived from the FrameNet and UD annotations, provided a manual mapping from lexical units in FrameNet to PropBank frames, and a mapping from FrameNet frame elements to PropBank semantic roles for the given pair of FrameNet and PropBank frames. Draft AMR graphs are to be derived from the UD and PropBank layers, as well as auxiliary layers containing named entity and coreference annotation, with the potential to seamlessly integrate the FrameNet frames and frame elements into the AMR graphs. The semantically richer FrameNet annotations (compared to PropBank) are also helpful in acquiring more accurate draft AMR graphs, even if FrameNet itself stays behind the scenes.

The inspiration to create an integrated multilayer corpus comes from the OntoNotes corpus (Hovy et al., 2006) and the Groningen Meaning Bank (GMB) (Bos et al., 2017). The overall difference from the OntoNotes approach is that we use the UD model at the treebank layer, and we annotate FrameNet frames in addition to the PropBank frames. In fact, FrameNet is the primary frame-semantic representation in our approach. Another difference is that we aim at whole-sentence semantic annotation at the ultimate AMR layer. This in some sense is similar to the goal of GMB, but GMB uses Discourse Representation Theory instead of AMR. For pragmatic reasons, we use the more shallow and more lossy AMR formalism. Our experience developing semantic parsers and multilingual text generators, by combining machine learning and grammar engineering (Gruzitis et al., 2017; Gruzitis and Dannells, 2017), has convinced us that FrameNet and AMR both have a great potential to establish as powerful and complementary semantic interlinguas which can be furthermore strengthened and complemented by other multilingual frameworks, like Grammatical Framework (Ranta, 2011).

In this paper, we focus on the creation of the intermediate FrameNet layer of the full-stack multilayer corpus. Note that the current project addresses only frequently used verbs as frame-evoking lexical units. A spin-off project has been just launched to work on frequent nominalizations, following the same methodology.

It should also be noted that there has been previous work on a domain-specific Latvian FrameNet for a real life media monitoring use case, focusing on 26 modified Berkeley FrameNet (BFN) frames (Barzdins et al., 2014). The current work, however, aims at a balanced general-purpose BFN-compliant framenet that will cover many frequently used frames and lexical units.

Although this paper focuses on Latvian, we believe that our experience and findings can be useful for the creation of dependency treebank based framenets for other less-resourced languages as well.

Table 1: A sample sentence *"On Wednesday evening, the nation's beloved poet Imants Ziedonis passed away at age 79."* represented in the CoNLL-U data format. Field FEATS is left empty because of space restrictions. The literal translations are not part of CoNLL-U.

ID	FORM	LEMMA		UPOSTAG	XPOSTAG	FEATS	HEAD	DEPREL	DEPS
1	Trešdienas	trešdiena	'Wednesday'	NOUN	ncfsg4		2	nmod	2:nmod:gen
2	vakarā	vakars	'evening'	NOUN	ncmsl1		7	obl	7:obl:loc
3	79	79	'79'	NUM	xn		4	nummod	4:nummod
4	gadu	gads	'year'	NOUN	ncmpg1		5	nmod	5:nmod:gen
5	vecumā	vecums	'age'	NOUN	ncmsl1		7	obl	7:obl:loc
6	mūžībā	mūžība	'eternity'	NOUN	ncfsl4		7	obl	7:obl:loc
7	aizgājis	aiziet	'leave'	VERB	vmnpdmsnasnpn		0	root	0:root
8	tautā	tauta	'nation'	NOUN	ncfsl4		9	obl	9:obl:loc
9	mīlētais	mīlēt	'love'	VERB	vmnpdmsnpsypn		10	amod	10:amod
10	dzejnieks	dzejnieks	'poet'	NOUN	ncmsn1		11	nmod	11:nmod
11	Imants	Imants	'Imants'	PROPN	npmsn1		7	nsubj	7:nsubj
12	Ziedonis	Ziedonis	'Ziedonis'	PROPN	npmsn2		11	flat:name	11:flat:name
13	.	.	'.'	PUNCT	zs		7	punct	7:punct

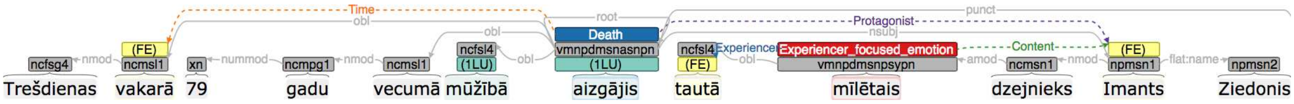


Figure 1: FrameNet annotation in WebAnno on top of a UD tree (Table 1). Only head nodes are selected while annotating frame elements (FE). The FE spans can be acquired automatically by traversing the respective subtrees: [*trešdienas vakarā*]$_{\text{Time}}$, [*tautā*]$_{\text{Experiencer}}$, [*tautā mīlētais dzejnieks Imants Ziedonis*]$_{\text{Protagonist}}$. Multi-word lexical units (LU) are indicated by generic LU tags: *mūžībā aizgājis*$_{\text{DEATH}}$ versus *mīlētais*$_{\text{EXPERIENCER\_FOCUSED\_EMOTION}}$.

2. The corpus

In this project, we are aiming at a medium-sized corpus – around 10,000 sentences annotated at all the layers mentioned in Section 1. Therefore it is crucially important to ensure that the multilayer corpus is balanced not only in terms of text genres and writing styles but also in terms of lexical units.

Our fundamental design decision is that the text unit is an isolated paragraph. The corpus therefore consists of manually selected paragraphs from many different texts of various types. Representative paragraphs are selected in different proportions from a balanced 10-million-word text corpus: around 60% come from various news sources, around 20% is fiction, around 10% are legal texts, around 5% is spoken language (parliament transcripts), and the rest is miscellaneous.

As for the lexical units, our goal is to cover at least 1,000 most frequently occurring verbs, calculated from the 10-million-word corpus. Since the most frequent verbs tend to be also the most polysemous, we expect that the number of lexical units (verb senses w.r.t. FrameNet frames) will be considerably larger – at least 1,500 units. At this stage, it is too early to predict any numbers regarding nominal lexical units. Nevertheless, the more frequent a lexical unit is, the more annotated examples it will have. We are aiming at around 10 annotation sets per lexical unit on average.

Paragraphs to be annotated are selected based on verbs they contain, not randomly, and curators are constantly updated on the current balance or imbalance of the corpus w.r.t. genres and verb frequencies. We assume that the corpus will turn out to be balances also w.r.t. nominal lexical units.

Our decision about the data selection is justified also by the lessons learned in other treebanking and framebanking projects. For instance, Bick (2017) concludes that a sentence-randomized propbank not only has a limited usage for coreference resolution and discourse analysis but also provides a limited coverage of lexical units.

At the time of writing, we have acquired more than 5,000 annotation sets (by investing around four man-months). This data set already covers more than 300 BFN frames evoked by nearly 900 lexical units.

The Latvian FrameNet corpus is being gradually released on GitHub under the CC BY-NC-SA 4.0 license.[1]

3. The FrameNet annotation process

Paragraphs for which the manual treebank annotation is finalized and which have been successfully converted to the UD representation are considered for the FrameNet annotation. Unfinished paragraphs are ignored till next iteration, since their sentence split, tokenization, as well as tree structure can still considerably change. Changes in the tree structure is not a major issue, and the FrameNet annotation process actually helps to spot and eliminate many inconsistencies in the underlying trees. The sentence splitting and tokenization, however, is a major requirement to later avoid issues in merging the different annotation layers.

Since we annotate FrameNet frames on top of UD trees, we need an annotation tool which supports both representations. Therefore we have chosen the WebAnno platform (Eckart de Castilho et al., 2016) which also supports a centralized web-based annotation workflow.

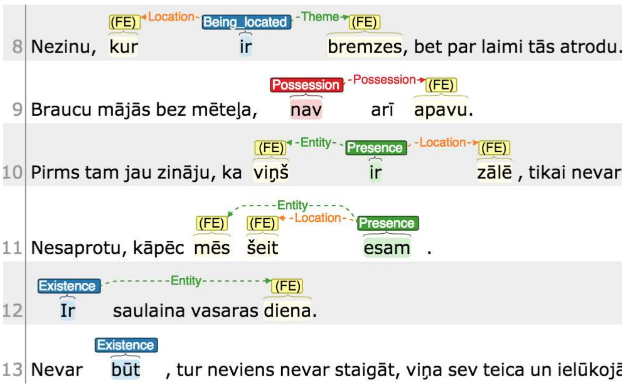
[1] https://github.com/LUMII-AILab/FullStack

3.1. The concordance approach

While treebank, named entity and coreference annotations are done paragraph by paragraph and sentence by sentence, we do not find this being a productive workflow for annotating semantic frames, especially in case of the highly abstract FrameNet frames. Instead, we prefer a concordance view, so that the linguist can focus on a target word and its different senses (frames), without constantly switching among different sets of frames. This also improves the annotation consistency.

To provide such annotation environment, we automatically extract all UD-annotated sentences from the finalized paragraphs containing the requested target word, and we store the result in a separate CoNLL-U file. More precisely, we group sentences for FrameNet annotation by applying filters on the LEMMA and POSTAG fields in the CoNLL-U files (see Table 1), as well as the DEPREL field in case of nominalizations (e.g. participles having the *amod* or *nmod* dependency relation).

Figure 2 illustrates a partial concordance with FrameNet annotations. The UD annotations are hidden for the sake of simplicity, and, in fact, they are hidden also in the curation view in WebAnno.[2] The actual annotation, however, is done on top of the UD layer, as illustrated in Figure 1.



8 Nezinu, kur ir bremzes, bet par laimi tās atrodu.

9 Braucu mājās bez mēteļa, nav arī apavu.

10 Pirms tam jau zināju, ka viņš ir zālē , tikai nevar

11 Nesaprotu, kāpēc mēs šeit esam .

12 Ir saulaina vasaras diena.

13 Nevar būt , tur neviens nevar staigāt, viņa sev teica un ielūkojā

Figure 2: A screenshot of the WebAnno tool: FrameNet-annotated occurrences of the target verb *būt* ('to be located', 'to be present', 'to have', or 'to exist').

When more paragraphs are finalized at the UD layer, they are included in the next concordance queries. In practice, for each target word there will be at least two concordance files extracted and annotated during the project. The first concordance is processed when there are at least three example sentences available for the target word. The second concordance will be processed when the planned 10,000 sentences will be done at the UD layer. The second concordance will contain only the new examples which are not included in the first concordance (according to the sentence identifiers). The annotated concordances from the first round will serve as guidelines when annotating the second round, thus, further improving consistency.

A consequence of such approach is that no full-text annotation is intentionally done, although many sentences might become fully or almost fully annotated after merging annotations of the same sentence from different concordances (see Table 2).

3.2. The UD-based annotation

The acquired UD-annotated concordances (full sentences) are imported in the WebAnno platform which we have specifically configured for the FrameNet annotation. To facilitate the annotation process, we have generated two kinds of WebAnno constraint sets. First, a set of frame to core frame element mappings (from BFN 1.7 data), so that a menu of core frame elements is generated when the annotator selects a frame for the particular occurrence of the target word. Second, a set of LEMMA/POSTAG to frame mappings, so that the most probable frames (senses) for the particular occurrence of the target word are highlighted at the top of the frame selection menu.

The UD-based approach has a significant consequence: frame elements are not annotated as spans of text – annotators select only the head word (node) when annotating a frame element. The whole span can be easily calculated automatically by traversing the respective UD subtree. These calculations are not included as part of the data set.

Such approach not only makes the annotation process more simple and the annotations more consistent, but it also facilitates the training of an automatic semantic role labeler, since it is easier to identify the syntactic head of a frame element than a span of a string. Still, most FrameNet corpora are annotated in terms of spans, relying on syntactic parsing as a post-processing step.

When the FrameNet annotation is done, the finalized concordances are exported from WebAnno, and are converted from the TSV3 format used by WebAnno to a more common CoNLL 2009-like format which combines the UD and FrameNet annotations (see Table 2). During conversion, the UD data fields in the CoNLL-2009 output are updated from the latest version of the UD treebank, and the isolated sentences are eventually reorganized back into paragraphs.

3.3. Important notes on frame elements

Yet another important decision regarding frame elements is to annotate only the core elements according to BFN. We have made this decision because of the limited time frame and the wider scope of the current project. However, we do annotate two non-core elements systematically: *Time* and *Place* (as illustrated in Figure 1). Our industrial partner, the national news agency LETA, is interested in the automation of media monitoring processes. In their information extraction use case, these two non-core frame elements are important, and they will be informative in other use cases as well. Other non-core elements are annotated occasionally, if they are rather specific to the frame (e.g. non-core indirect objects and specific adverbial modifiers).

Regarding null instantiations (NI), we do not annotate missing frame elements in the sentence. This is out of the scope of the current project, but the annotation of NI should to be considered in a follow-up research: (i) since the FrameNet annotation is relaying on UD, it is an open question how to handle NI – where to attach these annotations; (ii) since Latvian is a highly inflected language, the grammatical sub-

[2] Each concordance is annotated by one linguist and curated by another linguist, which is supported in WebAnno.

Table 2: A data format used to serialize the FrameNet layer of the corpus: a version of CoNLL-2009 based on CoNLL-U (see Table 1). Several CoNLL-U fields are excluded from this table because of space restrictions.

ID	FORM	LEMMA	UPOSTAG	XPOSTAG	DEPS	**FILLPRED**	**PRED**	**APRED$_1$**	**APRED$_2$**
1	Trešdienas	trešdiena	NOUN	ncfsg4	2:nmod:gen	_	_	_	_
2	vakarā	vakars	NOUN	ncmsl1	7:obl:loc	_	_	Time	_
3	79	79	NUM	xn	4:nummod	_	_	_	_
4	gadu	gads	NOUN	ncmpg1	5:nmod:gen	_	_	_	_
5	vecumā	vecums	NOUN	ncmsl1	7:obl:loc	_	_	_	_
6	mūžībā	mūžība	NOUN	ncfsl4	7:obl:loc	_	_	_	_
7	aizgājis	aiziet	VERB	vmnpdmsnasnpn	0:root	Y	Death	_	_
8	tautā	tauta	NOUN	ncfsl4	9:obl:loc	_	_	_	Experiencer
9	mīlētais	mīlēt	VERB	vmnpdmsnpsypn	10:amod	Y	Experiencer_focused_emotion	_	_
10	dzejnieks	dzejnieks	NOUN	ncmsn1	11:nmod	_	_	_	_
11	Imants	Imants	PROPN	npmsn1	7:nsubj	_	_	Protagonist	_
12	Ziedonis	Ziedonis	PROPN	npmsn2	11:flat:name	_	_	_	_
13	.	.	PUNCT	zs	7:punct	_	_	_	_

ject and object can be omitted in a sentence, to some extent, compensating it with the respective form of the verb; (iii) in general, it would require Latvian-specific guidelines, but the theoretical foundations are not mature yet for Latvian; it would require more elaborate linguistic research, based on the basic annotated data acquired in the current project; (iv) although NI is highly relevant for lexicographic research, it is not a priority for many practical use cases that require semantic parsing.

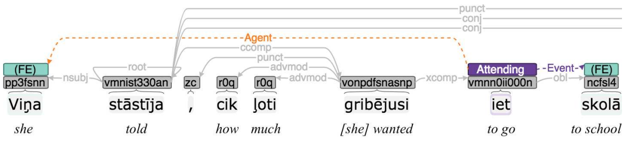


Figure 3: Non-projective annotation of a frame element (FE): the frame *Attending* is evoked in a subclause while its FE *Agent* is mentioned in the main clause.

It should be noted, however, that we do annotate frame elements that non-projective w.r.t. the underlying UD tree structure, i.e., that syntactically are not arguments of the target verb. Figure 3 provides an example.

3.4. Multi-word lexical units

Regarding lexical units, although we focus on verbs, they some times must be considered as multi-word units or constructions. To deal with this issue, we have introduced an auxiliary annotation layer for multi-word lexical units (as illustrated in Figure 1). The head word is still a verb that evokes a frame, but the other key constituents are indicated as well. Again, note that these constituents are roots of the respective subtrees (in general) – we do not annotate the whole spans.

This auxiliary layer is not an ultimate solution to deal with constructions, but for now it allows us to register such cases and to retrieve them later for more elaborated analysis. Usually these are partially grammaticalized constructions or even idioms that, as a whole, evoke the respective frames. If we would consider these verbs in isolation, they would rather evoke different frames, e.g.:

iet bojā 'to die' (*iet* – 'to go');

aiziet mūžībā 'to pass away' (*aiziet* – 'to leave');

ņemt vērā 'to consider' (*ņemt* – 'to take');

nākt klajā 'to be published' (*nākt* – 'to come');

nākt par labu 'to be beneficial' (*nākt* – 'to come');

likt lietā 'to use' (*likt* – 'to put').

3.5. Cross-lingual issues

In order to ensure compliance with the Berkeley FrameNet and, thus, to maximize the cross-lingual applicability of Latvian FrameNet, we are strictly sticking to the BFN frame inventory. We avoid defining any Latvian-specific frames. Therefore it is sometimes difficult to select an appropriate BFN frame for a particular sense of a Latvian verb. It usually happens when:

1. The sense of a Latvian verb is more specific compared to the closest English verb sense or compared to the definition of the closest BFN frame. For instance, for the verb *pārdomāt* 'to change one's mind' or 'to rethink', we do not have a solution yet, since BFN frames related to thinking (*Opinion*, *Cognitation*) do not fit this verb sense, and neither does the general *Cause_change* frame. Similarly, we have not found a good mapping for *maldīties* 'to be wrong' and *saņemties* 'to pull oneself together'.

2. The sense of a Latvian verb is more general compared to the closest English verb sense: the sense of an English verb is expressed in Latvian by a phrase (typically, by a verb and a direct object). Examples: *lasīt lekciju* 'to lecture' ('to give a lecture'), *krist ģībonī* 'to faint' ('to fall into unconsciousness'), *zaudēt samaņu* 'to faint' ('to lose consciousness').

3. The semantic elements are different between the Latvian and English verb senses. For instance, *braukt* 'to move using a vehicle': the sense of the Latvian verb does not specify whether the person is a driver or a passenger (e.g. *es braucu uz darbu* 'I go to work (by a transport)' – it is unclear what is the role of the person, and which frame is evoked – *Ride_vehicle* or *Operate_vehicle*. In this particular case, we use the frame *Use_vehicle* which is a non-lexical frame in English.

There are some options how to deal with these issues: (i) by treating more verb phrases in Latvian as if they were multi-word lexical units, even if lexicographers would argue about that (the second point in the above listing); (ii) by using a more general BFN frame if possible, i.e., if the direct object of the target verb can be annotated as a core frame element (e.g., it would work for ‘to lose consciousness’ but not for ‘to give a lecture’); (iii) some frames are just missing in BFN, and a global solution would be needed on how to propose and confirm new frames in the BFN frame hierarchy; most likely in the scope of the Multilingual FrameNet initiative (Gilardi and Baker, 2018).

4. Conclusion

Creating the Latvian FrameNet, we strictly follow a corpus-driven approach: no lexical units are introduced without annotated examples, i.e., we create no lexical units based on lexicographic intuition or a common-sense dictionary; only based on corpus evidence. An initial experiment on bootstrapping lexical units without corpus evidence did not prove to be productive: many of those hypothesis are not confirmed by our corpus (at least for now), and vice versa – many lexical units were missing.

The consecutive treebank and framebank annotation workflow has turned out very productive and mutually beneficial. The dependency tree facilitates the annotation of semantic frames and roles, while the frame semantic analysis of the verb valency often unveils various inconsistencies and bugs in the dependency or morphological annotation. These issues are immediately fixed in the treebank, and are later automatically synchronized with the FrameNet layer.

Because of the UD-based approach, we cannot use the specialized annotation tools developed for Berkeley FrameNet, or FrameNet Brasil, for instance. However, conversion to the BFN data format (from a CoNLL-like format) is possible (by using UD dependency relations instead of phrase types, etc.), so that the BFN-compliant web tools could be used at least for viewing and browsing Latvian FrameNet.

5. Acknowledgements

This work has received financial support from the European Regional Development Fund under the grant agreements No. 1.1.1.1/16/A/219 and No. 1.1.1.2/VIAA/1/16/188.

6. Bibliographical References

Banarescu, L., Bonial, C., Cai, S., Georgescu, M., Griffitt, K., Hermjakob, U., Knight, K., Koehn, P., Palmer, M., and Schneider, N. (2013). Abstract Meaning Representation for Sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria.

Barzdins, G., Gosko, D., Rituma, L., and Paikens, P. (2014). Using C5.0 and exhaustive search for boosting frame-semantic parsing accuracy. In *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC)*, pages 4476–4482, Reykjavik, Iceland.

Bick, E. (2017). From Treebank to Propbank: A Semantic-Role and VerbNet Corpus for Danish. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 202–210, Gothenburg, Sweden.

Bos, J., Basile, V., Evang, K., Venhuizen, N., and Bjerva, J. (2017). The Groningen Meaning Bank. In Nancy Ide et al., editors, *Handbook of Linguistic Annotation*, volume 2, pages 463–496. Springer.

Eckart de Castilho, R., Mújdricza-Maydt, E., Yimam, S. M., Hartmann, S., Gurevych, I., Frank, A., and Biemann, C. (2016). A Web-based Tool for the Integrated Annotation of Semantic and Syntactic Structures. In *Proceedings of the Workshop on Language Technology Resources and Tools for Digital Humanities*, pages 76–84, Osaka, Japan.

Fillmore, C. J., Johnson, C. R., and Petruck, M. R. (2003). Background to FrameNet. *International Journal of Lexicography*, 16(3):235–250.

Gilardi, L. and Baker, C. (2018). Learning to Align across Languages: Toward Multilingual FrameNet. In *International FrameNet Workshop 2018: Multilingual FrameNets and Constructicons*, Miyazaki, Japan.

Gruzitis, N. and Dannells, D. (2017). A multilingual FrameNet-based grammar and lexicon for Controlled Natural Language. *Language Resources and Evaluation*, 51(1):37–66.

Gruzitis, N., Gosko, D., and Barzdins, G. (2017). RIGOTRIO at SemEval-2017 Task 9: Combining machine learning and grammar engineering for AMR parsing and generation. In *Proceedings of the 11th International Workshop on Semantic Evaluation*, pages 924–928, Vancouver, Canada.

Gruzitis, N., Pretkalnina, L., Saulite, B., Rituma, L., Nespore-Berzkalne, G., Znotins, A., and Paikens, P. (2018). Creation of a Balanced State-of-the-Art Multilayer Corpus for NLU. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC)*, Miyazaki, Japan.

Hovy, E., Marcus, M., Palmer, M., Ramshaw, L., and Weischedel, R. (2006). OntoNotes: The 90% Solution. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 57–60, New York City, USA.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., Tsarfaty, R., and Zeman, D. (2016). Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, pages 1659–1666.

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

Pretkalnina, L., Rituma, L., and Saulite, B. (2016). Universal Dependency treebank for Latvian: A pilot. In *Human Language Technologies – The Baltic Perspective*, volume 289 of *Frontiers in Artificial Intelligence and Applications*, pages 136–143. IOS Press.

Ranta, A. (2011). *Grammatical Framework: Programming with Multilingual Grammars*. CSLI Publications, Stanford.

Deriving Mappings for FrameNet Construction from a Parsed Corpus of Japanese

Stephen Wright Horn, Alastair Butler, Iku Nagasaki, Kei Yoshimoto
NINJAL, Hirosaki University, NINJAL, Tohoku University
horn.s.w@ninjal.ac.jp, ajb129@hotmail.com, inaga@ninjal.ac.jp, kei@compling.jp

Abstract

A novel method is introduced for employing a SUSANNE / Penn Historical style parsed corpus to produce FrameNet mapping slots. Target/dependent pairings that are specified for a full range of basic grammatical relations can be generated. These can serve as slots for further specification as frame elements. As outcomes, gains in the speed and accuracy of FrameNet annotation are expected. The information about argument realisation patterns that a fully parsed corpus provides could be used as a basis to choose between senses of a given instantiation of word use, allowing for automated assistance with identifying frames and frame elements, given a sufficiently specified FrameNet. In a parsed corpus, basic local and non-local dependencies (argument/predicate, modifier/head, antecedent/pronoun, etc.) are exhaustively described by associating grammatical relations (dependencies) to specific tree structures. By meshing such an annotation schema with a semantic calculation, a rich range of dependencies can be established without recourse to overt indexing in the source annotation. The semantic calculation derives dependencies from the structure and records them in predicate logic formulas. These logical expressions can then be used to generate derived indices. The end result is a grammar-driven, exhaustive analysis of text into sets of target words and grammatically related dependents. Each set can serve as the co-domain for mapping FrameNet roles onto grammatical structures. The technique yields robust and flexible target/dependent pairings, and can be adapted to many different languages, illustrated here with an example from Japanese.

Keywords: Parsed Corpus, FrameNet, Source Annotation, Derived Annotation, Lexical Semantics

1. Introduction

This paper presents a novel way in which a SUSANNE / Penn Historical style parsed corpus (Sampson 1995, Santorini 2010) can be processed into a form enabling fast and accurate FrameNet annotation. The proposal is instantiated using the NINJAL Parsed Corpus of Modern Japanese (NPCMJ; NINJAL 2016). Corpora of this type define basic grammatical dependencies (argument/predicate, modifier/head, antecedent/pronoun, etc.) as relations within tree structures. Tree structures are defined by labelled nodes, and the relations of precedence and dominance that obtain between those nodes. Within this framework, annotators make exhaustive descriptions of argument realisation. Null elements are employed to mark up both local and non-local dependencies. Grammatical processes such as control, coordination, relativisation, and displacement are also defined. The innovation consists of an automatically derived intermediate analysis that expresses these relations with predicate logic based formulas (relations between predicate/variable bindings), derived as output from a Tarskian style semantic calculation (Tarski and Vaught 1956, Dekker 2012, Butler 2015) that makes reference to the grammatical analysis assigned in the parse.

The approach contrasts with methods that rely on indexing in the source annotation to express abstract relations between nodes, where an index might be a shared mark (e.g., a numeral), or might exist as a value (name) that references the position of an annotation component. Indexing for specifying relationships is applicable to data in any form, and is used in many annotation formats to specify the semantic roles that a constituent holds with respect to some lexical head. Building annotation with indexing is typically costly, in the sense that it often requires a human in the annotation chain to make the critical decision regarding the relation established. Unfortunately, when the namings for indices are motivated by reference to position in a sequence or structure, the dependencies those indices are used to establish can be easily broken by changes in orthography or segmentation, by the introduction of null elements, or by changes in structural assignment. Post-processing texts with such indices is also complicated: There is the need to preserve the motivation of the name of such an index together with the dependency it is meant to encode across changes in sequence or structure.

Our solution to these problems lies in pairing an annotation schema that encodes dependencies as relations in tree structures with a semantic calculation that re-expresses those dependencies as predicate logic bindings (Butler and Horn 2017). Obtaining semantic dependencies relies on the tree structure providing sufficient conditions for identifying grammatical relations, but the definitions of grammatical relations allow for a certain amount of variation in the position and make up of syntactic constituents. In a word, the basis for deriving dependencies is flexible, but the dependencies thus derived (expressed in a structure-independent format) are reliably robust.

The advantages of such a system are many, but its application in annotating semantic roles is particularly noteworthy. Derived indices can be generated from the semantic expressions in a text. These can be shared between a target and a particular dependent (mediated by the appropriate grammatical role) in a post-processing phase. An annotator can associate the appropriate semantic role (e.g., a frame-element in FrameNet annotation) with the index on the target, without the need to establish the pairing by hand. In this way the system supplies an objective and exhaustive basis for assigning semantic roles, where the human labour involved consists of filling an empty slot with a role name. In cases where FrameNet is able to make distinctions between word senses by reference to argument structure, the exhaustive description of an instantiation of word use in a parsed corpus could conceivably be used to automate the specification of a frame. More generally, embedding a FrameNet analysis in a fully developed description of discourse opens up new avenues of research, arguably multiplying the usefulness of both. For example, independently supported accounts of phenomena such as polysemy and construction meaning could be pursued in a corpus-based program of research.

The remainder of this paper is structured as follows. Section 2. outlines the principles by which basic grammatical relations are defined and shows their instantiation in an example. Section 3. shows how the index-less annotation is subsequently processed to enable a mechanism of semantic calculation that identifies dependencies by reference to structure, but re-expresses them in a structure-independent form. These dependencies can be subsequently assigned to structures through a derived indexing, but don't rely on indexing in order to be established. Section 4. outlines how the system can be harnessed for FrameNet annotation. Section 5. is a summary.

2. Parsed corpus annotation

We illustrate basic annotation principles using the following Japanese sentence:

(1)	ehon	o	kat	ta	kodomo	ga,	sore
	picture.book	ACC	buy	PAST	child	NOM	this

	o	oyatsu	o	tabe	nagara	yon	de	i	ta.
	ACC	snacks	ACC	eat	while	read	GER	exist	PAST

'The boy who bought the book was reading it while eating snacks.'

Local grammatical relations with respect to a predicate head are encoded through sisterhood under a clause node (IP) in conjuction with tag extensions for grammatical function ("`-SBJ`" for subject, "`-OB1`" for direct object, "`-ADV`" for adverbial, etc.). Consider these with respect to the verb *yon* 'eat' in the context of the matrix clause ("`IP-MAT`") of the annotation for (1):

(2)

```
( (IP-MAT (PP-SBJ (NP (IP-REL (NP-SBJ *T*)
                              (PP-OB1 (NP;{BOOK} (N ehon))
                                      (P-ROLE o))
                              (VB kat)
                              (AXD ta))
                      (N kodomo))
                  (P-ROLE ga))
          (PU ,)
          (PP-OB1 (NP;{BOOK} (PRO sore))
                  (P-ROLE o))
          (IP-ADV2-SCON (PP-OB1 (NP (N oyatsu))
                                (P-ROLE o))
                        (VB tabe)
                        (P-CONN nagara))
          (VB yon)
          (P-CONN de)
          (VB2 i)
          (AXD ta)
          (PU .))
  (ID example;JP))
```

Every basic grammatical function in the text in (1) is associated with a structural relation in the tree in (2). A relative clause is formed by a typed clause ("`IP-REL`") containing an index-less trace ("`(NP-SBJ *T*)`"). These are sufficient to associate the modified head *kodomo* 'child' with the subject argument for *kau* 'buy' in the relative clause by virtue of matching a generalized structural configuration on which the trace/relative head dependency is defined. The *nagara* 'while' clause is specified as subordinated ("`-SCON`") with a further specification ("`-ADV2`") requiring a subject-role argument as the antecedent for control (ruling out object *sore* 'this' as a potential antecedent). This is sufficient to establish the noun phrase headed by *kodomo* 'child' as controlling the index-less empty subject position for the verb *tabe* 'eat', again by virtue of matching a generalized structural configuration on which the subject control dependency is defined. Furthermore, sort information ("`;{BOOK}`") has been added to resolve the reference of the pronoun *sore* 'this' as co-valued with *ehon* 'picture.book'. In this way both local dependencies and non-local dependencies are established with a minimum of mark up language, and practically no recourse to overt indices in the annotation.

The above annotation schema is designed to be descriptively adequate for Japanese grammatical phenomena, but SUSANNE / Penn Historical style annotation can be adapted to describe many different languages. For each annotation schema a language-specific conversion can be applied to transform the data into a format that represents grammatical dependencies in a language-independent way.

3. Subsequent interpretation

This section sketches how structural relations are related to rules that interpret those relations as dependencies by a systematic conversion of the data. The conversion takes the form of a number of transformation steps, the first of which (tree normalisation) involves regularising tree structure and reducing the inventory of tag labels. Tag extensions are removed if redundant, or else can have their information contribution off-set. Also, particles marking core grammatical roles are substituted for the grammatical role they mark. Taking the Japanese tree in (2) as an example, the nominative case marker "`(P-ROLE ga)`" when under "`PP-SBJ`" is replaced by "`(P-ROLE ARG0)`", and "`-SBJ`" is removed. Other changes include collecting the verbal syntagm under one node with off-set information under an "`ACT`" node to preserve, e.g., tense information. In this way, the normalised tree in (3) is reached:

(3)

```
( (IP-MAT (ACT past)
          (PP (P-ROLE ARG0)
              (NP (CP-REL (IP-SUB (ACT past)
                                  (PP (P-ROLE ARG0)
                                      (NP *T*))
                                  (PP (P-ROLE ARG1)
                                      (NP (SORT *BOOK*)
                                          (N ehon)))
                                  (VB kat_ta)))
                  (N kodomo)))
          (PU ,)
          (PP (P-ROLE ARG1)
              (NP (SORT *BOOK*)
                  (PRO sore)))
          (PP (SORT *SITUATION*)
              (SCON *)
              (P nagara)
              (IP-ADV2 (PP (P-ROLE ARG1)
                           (NP (N oyatsu)))
                       (VB tabe)))
          (VB yon_de_i_ta)
          (PU .))
  (ID example;JP))
```

The second step is to convert the normalised tree into an expression that can serve as input to a semantic calculation system, specifically, the Scope Control Theory (SCT) system of Butler (2015). The normalized tree serves as input to a script that converts the data into an SCT expression, in which, for example, common nouns are treated as predicates taking "entity" variable bindings, verbs are treated as predicates taking "event" variable bindings, etc. Structure in SCT expressions is built exploiting normalized tree structure by locating any complement for the phrase head to scope over, adding modifiers as elements that scope above the head, and keeping track of the binding names (e.g., ""ARG0"" (logical subject)) for the resulting SCT expression. Conversion adds construction information from the constituent nodes (e.g, "subord" (subordinate clause), and "Lam ("h", "T", ...)" (an instruction to make the open ""h"" binding (the head binding internal to a noun phrase) into a ""T"" binding (the trace binding internal to a relative clause)), etc.). Conversion also adds instructions (e.g., "gen "EVENT"") to generate what will become bound variables of a resulting semantic calculation.

The overall output from conversion to an SCT expression for (1) is in (4) below:

(4)

```
val sent =
( fn fh =>
  ( fn lc =>
    ( ( fn lc =>
        ( some lc fh ".e" ( gen "ENTITY")
          ( scon fh "&"
            ( Lam ( "h", "T",
                subord lc nil
                ( ( fn lc =>
                    ( ( arg "T") "ARG0"
                      ( ( fn lc =>
                          ( some lc fh ".e" ( gen "BOOK")
                            ( nn lc "ehon")))
                        [ "ARG0", "ARG1", "h"] "ARG1"
                        ( past ".event"
                          ( verb lc ".event"
                            ["ARG1", "ARG0"] "kat_ta"
                            ( gen "EVENT"))))))
                  [ "ARG0", "ARG1"])))
            ( nn lc "kodomo"))))
      [ "ARG0", "ARG1", "h"] "ARG0"
      ( ( pro ["*"] [ "BOOK"] ".e" "sore" ( gen "BOOK")) "ARG1"
        ( scon fh "SCON_nagara"
          ( control2 lc
            ( ( fn lc =>
                ( ( fn lc =>
                    ( some lc fh ".e" ( gen "ENTITY")
                      ( nn lc "oyatsu")))
                  [ "ARG0", "ARG1", "h"] "ARG1"
                  ( verb lc ".event" ["ARG1"] "tabe"
                    ( gen "EVENT"))))
              [ "ARG0", "ARG1"]))
          ( past ".event"
            ( verb lc ".event" ["ARG1", "ARG0"] "yon_de_i_ta"
              ( gen "EVENT")))))))
  [ "ARG0", "ARG1"])
[ ".e", ".event"]
```

Following an evaluation of the above SCT expression, the predicate logic representation with sorted variables in (5) below is returned:

(5)

```
exists BOOK[4] EVENT[3] EVENT[6] EVENT[7] BOOK[2] ENTITY[5]
 ENTITY[1].(
    ehon(BOOK[2]) & kat_ta(EVENT[3],ENTITY[1],BOOK[2])
  & kodomo(ENTITY[1]) & BOOK[4] = BOOK[2]
  & oyatsu(ENTITY[5]) & past(EVENT[3])
  & past(EVENT[7]) & (tabe(EVENT[6],ENTITY[1],ENTITY[5])
    SCON_nagara yon_de_i_ta(EVENT[7],ENTITY[1],BOOK[4])))
```

Note how the predicate logic representation expresses the subjecthood of *kodomo* ("ENTITY[1]") and the objecthood of *ehon* ("BOOK[2]") and the action of buying ("EVENT[3]") as variables bound by the predicate kat_ta, even though *kodomo* does not appear locally with the verb. Identity between the book (*ehon*) that was bought and the pronoun (*sore*) standing in for the thing that was read is expressed as equality between two variables: "BOOK[4] = BOOK[2]". And control from the subject of the upstairs verb yon_de_i_ta (*was reading*) into the clause headed by tabe (*eating*) is captured by the way that "ENTITY[1]" is a bound argument of both predicates.

To illustrate the flexibility of the combination of structural definitions for grammatical relations and their rendering into predicate logic representations, consider a frame in which the verb *kau* is used in a sense including a beneficiary (e.g., *Boku wa otooto ni ehon o katta* "I bought a book for my little brother"). Recognising this sense, an annotator adds a null indirect object pronoun (NP-OB2 *pro*) as the first constituent in the relative clause in (2), thereby shifting the position of every other element in the tree. Notwithstanding, the definitions for subject, object, etc., still obtain, and these dependencies remain intact.

So far we have seen the automatic calculation of dependencies through structural assignments and their expression in structure-independent representations. All relations expressible in a well-formed parse are defined under the calculation. The accuracy of the calculation is directly related to the accuracy of the sourced parsed annotation. Using another automated process, a derived analysis such as that in (5) can be embedded back into the source phrase structure tree annotation to yield the tree in (6) below:

(6)

```
( (IP-MAT;@0:66
    (PP-SBJ;<0:22>;@0:22
      (NP;@0:19
        (IP-REL;<0:12>;@0:12
          (PP-OB1;<0:5>;@0:5 (NP;{BOOK};@0:3 (N;@0:3 ehon))
                             (P-ROLE;@5:5 o))
          (VB;<,0:5@ARG1,14:19@ARG0,EVENT[3]@EVENT,>;@7:9 kat)
          (AXD;@11:12 ta))
        (N;<,0:12@REL,0:22@h,>;<14:19>;@14:19 kodomo))
      (P-ROLE;@21:22 ga))
    (PU;@24:24 ,)
    (PP-OB1;<26:31>;@26:31
      (NP;{,0:5,};{BOOK};@26:29 (PRO;@26:29 sore))
                                (P-ROLE;@31:31 o))
    (IP-ADV2-SCON;@33:52
      (PP-OB1;<33:40>;@33:40 (NP;@33:38 (N;@33:38 oyatsu))
                             (P-ROLE;@40:40 o))
      (VB;<,33:40@ARG1,0:22@ARG0,EVENT[7]@EVENT,>;@42:45 tabe)
      (P-CONN;@47:52 nagara))
    (VB;<,SITUATION[5]@LINK,26:31@ARG1,0:22@ARG0,
          EVENT[8]@EVENT,>;@54:56 yon)
    (P-CONN;@58:59 de)
    (VB2;@61:61 i)
    (AXD;@63:64 ta)
    (PU;@66:66 .))
  (ID example;JP))
```

This “indexed” view gives a view of the tree structure with nodes dominating a constituent given a suffix “`@n:m`” where `n` is the `n`-th character of the overall tree yield (the collected terminal strings (words) retaining linear ordering and with word separations counting as single characters) marking the first character resulting from a yield of the constituent, while `m` is the `m`-th character of the tree yield marking the last character resulting from a yield of the constituent. In addition, indexing information is given to specify argument relationships and antecedence relationships having the form “`m:n`”, with the same ‘yield-span’ use of `n` and `m` as just described. The indexing information gives explicit indexing of grammatical dependencies that the original annotation had left implicit. Specifically the indexing makes the following contributions:

- Indexing given the form “`<n:m>`” marks a yield-span that serves as an argument for a predicate, as well as providing an antecedent for anaphoric reference.

- The arguments that a predicate takes are marked on the pre-terminal node for the predicate with a “`<,...,n:m@role,..,>`” format, with “`n:m`” providing information to locate the argument and “`role`” stating the argument role.

- Pronominal information is presented with the format “`{,...,n:m,...,}`”, that is, specifying potentially multiple antecedents.

Also note how the trace “`(NP-SBJ *T*)`”, as a zero element that would merely duplicate information now captured by the indexing, has been removed from the tree, thereby removing it from the calculation of the tree yield. With the targets for dependencies spelled out in the predicate nodes, this is now a basis for deriving as output the kind of formatted annotation seen with FrameNet.

4. FrameNet annotation

The immediate utility of this approach is apparent when one considers that FrameNet generalizes over collocations in which dependents are normally related to their targets through grammatical relations. FrameNet (Ruppenhofer et al 2016) uses role labeling annotation to anchor semantic frames to instantiations in natural language. In FrameNet annotation, a frame-element relates to a frame that subsumes multiple predicates with various manifestations for frame-specific semantic roles. Predicates and their arguments are anchored to the character string of the source data through a FrameNet report. To demonstrate how target/dependent pairs identified by a semantic calculation on the grammar can potentially be transformed into frame/frame-element pairs in the FrameNet XML annotation format, consider (7) below, which is the output from the pipeline described here, given the data in (1) as input. (7) is an underspecified FrameNet report generated directly from the tree in (6) with yield-span index information. The FrameNet information that remains to be added is represented by attributes with numbered blanks: “`attribute="_n_"`”.

(7)

```
<sentence>
  <text>ehon o kat ta kodomo ga , sore o oyatsu o tabe
      nagara yon de i ta .</text>
  <annotationSet luID="_1_" luName="kat.v" frameID="_2_"
               frameName="_3_">
    <layer rank="1" name="Target">
      <label end="9" start="7" name="_4_"/>
    </layer>
    <layer rank="1" name="FE">
      <label end="19" start="14" name="_5_"/>
      <label end="5" start="0" name="_6_"/>
    </layer>
  </annotationSet>
  <annotationSet luID="_7_" luName="kodomo.n" frameID="_8_"
               frameName="_9_">
    <layer rank="1" name="Target">
      <label end="19" start="14" name="_10_"/>
    </layer>
    <layer rank="1" name="FE">
      <label end="12" start="0" name="_11_"/>
    </layer>
  </annotationSet>
  <annotationSet luID="_12_" luName="tabe.v" frameID="_13_"
               frameName="_14_">
    <layer rank="1" name="Target">
      <label end="45" start="42" name="_15_"/>
    </layer>
    <layer rank="1" name="FE">
      <label end="22" start="0" name="_16_"/>
      <label end="40" start="33" name="_17_"/>
    </layer>
  </annotationSet>
  <annotationSet luID="_18_" luName="yon.v" frameID="_19_"
               frameName="_20_">
    <layer rank="1" name="Target">
      <label end="56" start="54" name="_21_"/>
    </layer>
    <layer rank="1" name="FE">
      <label end="22" start="0" name="_22_"/>
      <label end="31" start="26" name="_23_"/>
    </layer>
  </annotationSet>
</sentence>
```

To spell it out in more detail, the FrameNet format consists of source character data as the `<text>` content, followed by annotations for the predicates as `<annotationSet>` content. Predicates are picked out with `start` and `end` attributes for a `Target`. For example, the Target for the annotationSet with “`luName="kat.v"`” is the 8th (“`start="7"`”) to 10th (“`end="9"`”) characters of the text content, namely, “`kat`”. Arguments are similarly established as spans of characters of the source string. For example, the element corresponding to the character span `kodomo` (identified by “`start="14"`” and “`end="19"`”) is specified as having a role with respect to “`luName="kat.v"`”. The role would fill the blank in “`name="_5_"`”.

The annotation method we propose involves adding FrameNet information to target/dependent sets in the tree structures themselves. Such information can be added as the terminal strings of offset nodes which are adjacent to the target element and include information to pinpoint the ID of the relevant lexical unit—from which all frame details are recoverable—as well as the frame-elements applicable to target/dependent sets linked to the tree structure via the relevant grammatical function information. Note, for example, that a blank “`*_5_*`” appears with an “`ARG0`” node, which corresponds to a subject grammatical role in the tree (8) below.

(8)

```
( (IP-MAT (PP-SBJ (NP (IP-REL (NP-SBJ *T*)
                              (PP-OB1 (NP;{BOOK} (N ehon))
                                      (P-ROLE o))
                              (VB kat)
                              (FRAME (LU *_1_*)
                                     (ARG0 *_5_*)
                                     (ARG1 *_6_*))
                              (AXD ta))
                      (N kodomo))
                  (P-ROLE ga))
          (PU ,)
          (PP-OB1 (NP;{BOOK} (PRO sore))
                  (P-ROLE o))
          (IP-ADV2-SCON (PP-OB1 (NP (N oyatsu))
                                (P-ROLE o))
                        (VB tabe)
                        (FRAME (LU *_12_*)
                               (ARG0 *_16_*)
                               (ARG1 *_17_*))
                        (P-CONN nagara))
          (VB yon)
          (FRAME (LU *_18_*)
                 (ARG0 *_22_*)
                 (ARG1 *_23_*))
          (P-CONN de)
          (VB2 i)
          (AXD ta)
          (PU .))
  (ID example;JP))
```

A frame-element value added to the blank "*_5_*" fills the appropriate place in an output transformed to FrameNet format. Given tree structures with frame information added *in situ*, completed FrameNet reports could be produced automatically. The immediate benefits include being able to refer to an exhaustive grammatical analysis in the process of assigning frame-elements, and being able to take advantage of robust dependent/target links that have been established in advance. The ability to make changes in structural assignment and segmentation (within the parameters of the definitions in the syntactic annotation) without the danger of interrupting dependencies is an added advantage.

Using a parsed corpus as source data for FrameNet annotation increases in value proportional to the richness and accuracy of the source parse. One precedent for such an undertaking is the SALSA project (Burchardt et al 2006). Embedding FrameNet information into a well-articulated description of discourse could open up new avenues for research. We propose that such an undertaking would also enjoy increased productivity and accuracy if integrated into systems such as those being developed in tandem with SCT. Extending the application, argument structure profiles for specific instantiations of words in the parsed data could be used to filter appropriate candidates for frame assignments, further reducing the burden on human annotators.

5. Summary

To sum up, the proposal is to take advantage of a technique of annotation that shifts the role of indexing onto the assignment of structural positions in a syntactic tree, and supplies an interpretive process that creates the specifications of dependencies. Language specific source annotation is expressed in a language independent form as the result of a semantic calculation. So far the system has been applied for obtaining valence frames in English[1], Contemporary Japanese[2], and Old Japanese[3]. The system can be used to generate sets of target/dependent pairings that directly correspond to sets of frame/frame-element pairings in FrameNet analyses. We show how assignments of values for frame/frame-element pairings can be added directly to tree structures, and how the resulting tree structures can be processed to give outputs as FrameNet reports.

Acknowledgements

This paper describes work funded as a component of the NINJAL Parsed Corpus of Modern Japanese. We thank all project members for their help and advice. This paper benefitted from the comments of three anonymous reviewers, who we gratefully acknowledge.

6. Bibliographical References

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

A. Butler. 2015. *Linguistic Expressions and Semantic Processing: A Practical Approach*. Heidelberg: Springer-Verlag.

A. Butler and S.W. Horn. 2017. Annotating syntax and lexical semantics with(out) indexing. In *Proceedings of Logic and Engineering of Natural Language Semantics 14 (LENLS 14)*, Paper 24, pp. 1–12. University of Tsukuba: JSAI-isAI 2017.

P. Dekker. 2012. *Dynamic Semantics*, vol. 91 of *Studies in Linguistics and Philosophy*. Dordrecht: Springer Verlag.

NINJAL. 2016. NINJAL Parsed Corpus of Modern Japanese. (Version 1.0). National Institute for Japanese Language and Linguistics. (http://npcmj.ninjal.ac.jp/interfaces/ accessed 2018/02/28).

J. Ruppenhofer, M. Ellsworth, M.R.L. Petruck, C.R. Johnson, C.F. Baker, and J. Scheffczyk. 2016. FrameNet II: Extended Theory and Practice. Tech. rep., Berkeley.

G.R. Sampson. 1995. *English for the Computer: The SUSANNE Corpus and Analytic Scheme*. Oxford: Clarendon Press (Oxford University Press).

B. Santorini. 2010. Annotation manual for the Penn Historical Corpora and the PCEEC (Release 2). Department of Computer and Information Science, University of Pennsylvania, Philadelphia. (http://www.ling.upenn.edu/histcorpora/annotation).

A. Tarski and R.L. Vaught. 1956. Arithmetical extensions of relational systems. *Compositio Mathematica* 13:81–102.

[1] http://www.compling.jp/ajb129/tspc.html

[2] http://npcmj.ninjal.ac.jp/interfaces/index_en.html

[3] http://www.compling.jp/m97pc

Uneek: a Web Tool for Comparative Analysis of Annotated Texts

Per Malm[1], **Malin Ahlberg**[2], **Dan Rosén**[2]
[1]Department of Scandinavian Languages, Uppsala University, Sweden
[2]Språkbanken, University of Gothenburg, Sweden
per.malm@nordiska.uu.se, malin.ahlberg@gu.se, dan.rosen@svenska.gu.se

Abstract

In this paper, we present *Uneek*, a web based linguistic tool that performs set comparison operations on raw or annotated texts. The tool may be used for automatic distributional analysis, and for disambiguating polysemy with a method that we refer to as *semi-automatic uniqueness differentiation* (SUDi). Uneek outputs the intersection and differences between their listed attributes, e.g. POS, dependencies, word forms, frame elements. This makes it an ideal supplement to methods for lumping or splitting in frame development processes. In order to make some of Uneek's functions more clear, we employ SUDi on a small data set containing the polysemous verb *bake*. As of now, Uneek may only run two files at a time, but there are plans to develop the tool so that it may simultaneously operate on multiple files. Finally, we relate the developmental plans for added functionality, to how such functions may support FrameNet work in the future.

Keywords: frame development, distributional method, automated comparative analysis, polysemy disambiguation

1. Introduction

Uneek is a web based linguistic tool that automatizes complex comparative tasks. It takes two input files (txt or xml) and outputs their intersection, and/or the differences between them. This makes Uneek a suitable aid in frame development processes, e.g. to the splitting or lumping schemes described in Ruppenhofer et al. (2016). The benefits of the program is further illustrated in an example of polysemy disambiguation through a method we refer to as *semi-automatic uniqueness differentiation* (SUDi).

There are other approaches available to polysemy disambiguation. For example, one may choose a more statistical approach such as Drouin (2003) using *TermoStat*, a software designed for term extraction that determines the specificity of words in a domain-specific corpus compared to a larger reference corpus. One may also choose a more qualitative approach, e.g. Ruppenhofer et al. (2016) where the disambiguation of a polysemous form is based on the semantic frames they evoke. In this setting, SUDi may be considered a supplementary method to the same problem.

The paper is organized as follows: in Section 2, we present Uneek. Section 3 holds a presentation on how to use Uneek for polysemy disambiguation in SUDi. The final Section 4 contains some closing remarks and plans for future work.

2. Uneek

Uneek is an open source project, and the code is available under the MIT-license.[1] It is a tool for automatically performing distributional analyses in the sense of Harris (1954), where the "distribution of an element will be understood as the sum of all its environments". There are other programs available today that gives a similar result, e.g. *AntConc* (Anthony, 2016) and *Wordsmith* (Scott, 2017). One downside with the former is that it – to our best knowledge – is not currently designed to handle xml tags.[2] Consequently, it only operates on word level. One downside with the latter is that it is developed for Windows OS, and is not compatible with all operating systems. Uneek handles both txt and xml, and is available for online use by any modern web browser without specific OS requirements.

The chief benefit of Uneek lies in its ability to operate on annotated text. There are a number of freely available tools for automatic annotation, e.g. *Sparv*, an easy to use annotation pipeline for various languages (Borin et al., 2016), and *Stanford CoreNLP* (Manning et al., 2014).[3] Uneek may also be used on annotations from the Berkeley FrameNet.[4] Working on the level of annotations also gives the opportunity to compare texts in different languages, given that they have at least one annotation layer in common. This might be of interest when working with language independent frameworks, such as UD (Nivre et al., 2016).

Uneek has three general settings for (i) set comparison operations, (ii) input format, and (iii) shallow syntactic sequencing. These settings are presented in detail below.

There are two set comparison operations, the intersection ($A \bigcap B$), which we refer to as *intersectional analysis*, and the differences ($A - B$ and $B - A$) which we refer to as *uniqueness differentiation*. Uniqueness differentiation is used for SUDi, or other methods where a full account of the differences between two sets is wanted. For instance, consider the sets A and B in example 1a–b below.

(1) a. $A = \{Aegon, forgave, his, goat\}$
b. $B = \{Aegon, hid, his, goat, yesterday\}$

A uniqueness differentiation of the sets in example 1a–b results in the following two sets:

(2) a. $A - B = \{forgave\}$
b. $B - A = \{hid, yesterday\}$

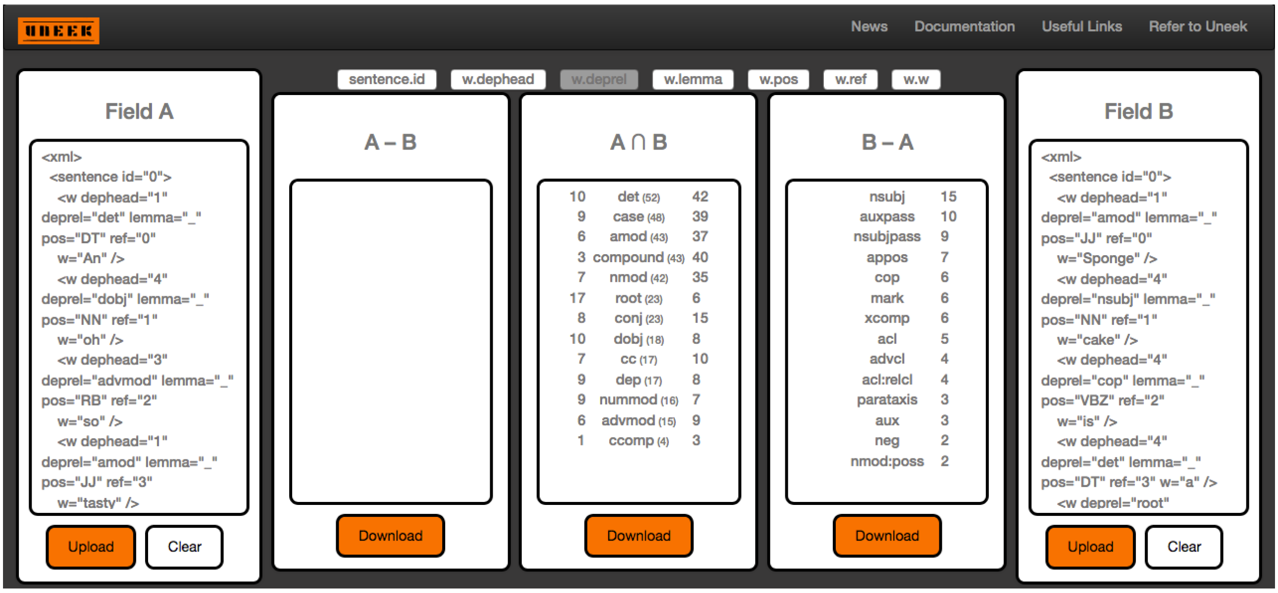
The intersectional analysis may be used for cases where a full account of what the two sets have in common is wanted. It provides the following set:

[1] The code is found at https://github.com/PerMalm/uneek, and the tool at https://uneek-tools.github.io/.

[2] However, this feature is under development: [last checked 11-01-2018] http://www.laurenceanthony.net/software/antconc/.

[3] There are other tools for FN annotation. See *SEMAFOR* for automatic annotation (Das et al., 2010), and *FrameNet Brasil WebTool* for manual annotation (Torrent et al., 2018).

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



Field A

```
<xml>
 <sentence id="0">
  <w dephead="1"
deprel="det" lemma="_"
pos="DT" ref="0"
  w="An" />
  <w dephead="4"
deprel="dobj" lemma="_"
pos="NN" ref="1"
  w="oh" />
  <w dephead="3"
deprel="advmod" lemma="_"
pos="RB" ref="2"
  w="so" />
  <w dephead="1"
deprel="amod" lemma="_"
pos="JJ" ref="3"
  w="tasty" />
```

A ∩ B

10	det (52)	42
9	case (48)	39
6	amod (43)	37
3	compound (43)	40
7	nmod (42)	35
17	root (23)	6
8	conj (23)	15
10	dobj (18)	8
7	cc (17)	10
9	dep (17)	8
9	nummod (16)	7
6	advmod (15)	9
1	ccomp (4)	3

B – A

nsubj	15
auxpass	10
nsubjpass	9
appos	7
cop	6
mark	6
xcomp	6
acl	5
advcl	4
acl:relcl	4
parataxis	3
aux	3
neg	2
nmod:poss	2

Field B

```
<xml>
 <sentence id="0">
  <w dephead="1"
deprel="amod" lemma="_"
pos="JJ" ref="0"
  w="Sponge" />
  <w dephead="4"
deprel="nsubj" lemma="_"
pos="NN" ref="1"
  w="cake" />
  <w dephead="4"
deprel="cop" lemma="_"
pos="VBZ" ref="2"
  w="is" />
  <w dephead="4"
deprel="det" lemma="_"
pos="DT" ref="3" w="a" />
  <w deprel="root"
```

Figure 1: A Uneek analysis of *a recipe for a sponge cake* (Field A) and *a description of sponge cake* (Field B)

(3) $A \bigcap B = \{Aegon, his, goat\}$

Even though these operations are simple, it is helpful to calculate them automatically. To manually perform these tasks on large texts, is both time consuming and error prone. The user chooses between two input formats. If *txt* is chosen, i.e. raw text, a simple whitespace tokenizer is run. The tokenized word forms are then given as input to the set comparisons. If xml format is chosen, all text nodes (typically word forms) as well as all attributes of all tags (typically annotations) are given as input.

A screenshot of Uneek is presented in Figure 1. In the leftmost box labeled Field A, the first input file is uploaded, and the second file is put in the rightmost box (Field B). In this particular case, we have set Uneek to perform an intersectional analysis and a uniqueness differentiation, and uploaded a recipe for sponge cake (File A), and some general description about sponge cake (File B). The input files have been processed using the Stanford Dependency Parser.[5] The result is presented in the three middle boxes. The second rightmost box and the second leftmost box holds the result of the uniqueness differentiation. The middle box shows the result of the intersectional analysis.

The attributes of the xml – corresponding to annotation layers – are visualized as radio buttons above the result boxes. These buttons control which layer is shown in the result box. In Figure 1 we have chosen to look at dependencies, which among other things tells us that the recipe for sponge cake lacks nominal subjects (nsubj). This is to be expected due to the imperative mood of recipes.

The third general setting allows for set comparisons on shallow syntactic sequences. A shallow syntactic sequence is here understood as a left to right organization of linguistic units specified in the xml. The only assumption made for the shallow syntax function is that the xml tag `sentence` sets the span in which the syntax chains are constructed. Below, we show some syntactic sequences in various annotation layers for example 1a.

(4) Text $= \{Aegon, forgave, his, goat\}$
Dep $= \{$nsubj, root, nmod:poss, dobj$\}$
Frame $= \{$Forgiveness$\}$
Frame Element $= \{$Judge, Evaluee, [INI]$\}$

The application of set operations on the syntactic sequences returns all the unique syntactic configurations for File A and File B, and all of their shared configurations. This function is especially useful for lexico-grammatical purposes. One may quite easily get a complete account of all the combinatorial possibilities of surface forms for complex constructions and frames. For instance, this function would probably ease the lexicographic frame annotation mode mentioned in Ruppenhofer et al. (2016, 19) by automatically listing all combinations of frame elements (FEs). To minimize visual clutter, the GUI only provides descriptive statistics (raw numbers). The output data may be downloaded in a human and machine readable format (csv) to ease export to statistical programs.

In sum: Uneek operates on user defined data, either raw or annotated text, and provides formal support for intuitions on lumping or splitting linguistic units. It may be used for automatic distributional analysis or for the disambiguation of polysemy presented next.

3. Semi-automatic Uniqueness Differentiation

The rationale for SUDi and Uneek rest on the distributional hypothesis (Harris, 1954) and set theory; see *locus classicus* Cantor (1915). Regarding the former, Firth (1962) wrote, ”You shall know a word by the company it keeps”. However, for the treatment of polysemy we assume: you shall know the difference between two polysemous words by the company one of them constantly rejects. In this section, we present the details of our proposed method for pol-

[5] StanfordDependencyParser from nltk.parse.stanford, v. 3.2.4.

ysemy disambiguation, called semi-automatic uniqueness differentiation (SUDi).

As we indicated in the previous section, there are two reasons for developing Uneek and SUDi. The first is to facilitate finding formal support for linguistic intuitions in complex material. The second is to improve the reliability of the distributional analysis. SUDi is a formalized and semi-automized methodology of some of the work that linguists often do: collect data, sort it and look for differences.

Roughly, SUDi involves five steps: (i) collect cases containing the presumed polysemous form from a corpus, (ii) sort these intuitively into two text-files (iii) process the files using an annotation device that produces xml which is needed in the next step, (iv) run the xml-files through Uneek, and (v) interpret the result.

Using Uneek in step (iv) not only speeds things up, but also simplifies reproducibility. However, to ensure validity, the user should (among other things) delimit the specified environment for the polysemous form in the input data. As for any tool, garbage in results in garbage out.

If Uneek does not find unique forms for one of the files, there is no formal support for polysemy. But, if it does, a linguist needs to interpret the result.

Step (v) in SUDi is based on proof by contradiction using human grammaticality judgements. First, take the linguistic unit that is unique in one of the files, and place it in the context of the polysemous item in the other file. If this switch leads to a semantic change that is deemed unfit in the tested domain (here marked with #), then there is *positive* formal support to the intuition that the polysemous form may be split into different frames, constructions, and so on. Though, if the linguistic unit works fine in the other context, then there is *negative* formal support for polysemy. Being unique in one domain does not lead to its infelicity in the other; uniqueness must be validated. Let us illustrate this step with an example case for which we strongly expect positive support for polysemy, namely for the verb *bake*.

First, we collect example sentences for *bake* from the Berkeley FrameNet Cooking creation frame and for *bake* from the Apply heat frame.[6] Second, we sort these in two files. Third, we automatically process them (again using the Stanford Dependency Parser). Fourth, we run the files through Uneek, and interpret the unique differences using the method in step five. The result of step four is presented in Table 1 where only some of the unique values for the Cooking creation *bake.v* are given.

Table 1: Unique values for *bake* (Cooking creation)

attributes	values (in absolute numbers)
Dep-heads:	auxiliary (10), predeterminer (2)
POS:	modal verb (5), poss. pronoun (4)
Words:	*Sunday* (2), *cakes* (1), *Saturday* (1)

Recall the uniqueness differentiation between the description and the recipe of sponge cake. Here we expect similar results to support that the difference between the Cooking creation and the Apply heat frames, lies in the latter being more recipe-like. For instance, the unique distribution with auxiliaries and modals in Table 1 is explained by the fact that recipes are written in the imperative mood.

Next, we test some of the unique values in Table 1 against a Berkeley FrameNet example from the Apply heat frame, i.e. *Bake the soufflés for 12 minutes*. These tests are presented in example 5a–d below.

(5) a. # Bake *all* the soufflés for 12 minutes.

b. # Bake *your* soufflés for 12 minutes.

c. Bake the soufflés $\left\{\begin{array}{l}\text{\# on \textit{Saturday}}\\ \text{for 12 minutes}\end{array}\right\}$.

d. Bake the *cakes* for 12 minutes.

From example 5a, we notice that *all* does not fit very well; it may be hard to find recipes for multi-soufflé cooking. Another rare bird in the recipe genre is to state the owner of the soufflé, as in example 5b, so is the instruction of cooking on specific weekdays (example 5c). On the contrary, these words work well in the Cooking creation frame, e.g. *Don't worry darling! I'll bake all your soufflés tomorrow.* However, observe that the unique form *cakes* (example 5d) can occur in the Apply heat frame. Hence, it is important to manually interpret the unique units, especially with the open word classes being what they are, i.e. open.

As a corroborative digression, we apply SUDi on the FEs in the annotated sentences for the Cooking creation and the Apply heat frame. The result is shown in Table 2.

Table 2: Unique and shared frame elements for *bake* in the Cooking creation (A) and the Apply heat frame (B)

$A - B$	$A \bigcap B$	$B - A$
Ingredients, Place, Recipient, Time, Produced food, Purpose	Target Cook Food	Temperature setting, Heating instrument, Duration, Manner, Container

Among other things, we note that the unique FEs Recipient and Time support the observations that were based on example 5b–c above. In conclusion: auxiliaries, possessives, and predeterminers indicate positive formal support for polysemy. But keep in mind the scarce input (36 sentences).

Speed and reliability of automized linguistic labour must not come at any cost, especially not at the price of validity. One should think twice before taking the human element out of the equation. A similar point is made in Fillmore (1992) about the pitfall of exclusively relying on intuitive data or empirical data, a point he makes clear by the following interaction between two radicalized linguists:

> [. . .] the corpus linguist says to the armchair linguist, ”Why should I think that what you tell me is true?”, and the armchair linguist says to the corpus linguist, ”Why should I think that what you tell me is interesting?” Fillmore (1992)

[6] https://framenet.icsi.berkeley.edu/fndrupal/

Fillmore argues for the need of both these radicals, i.e. a computer aided armchair linguist, who checks his/her grammaticality judgements against a corpus in some organized fashion. Still, even behind results that are ever so true and interesting, methodological problems may sometimes lurk about unnoticed, especially in manually performed distributional analyses. When faced with such cases, it is sensible to ask: why should I think that what you tell me is based on a reliable method? A true and interesting result does not necessarily paint the whole distributional picture. We want to be able to say that given a specific corpus and a specific method, we will always get the complete distribution of a particular linguistic unit. We believe that Uneek and SUDi allows linguists to make such statements.

4. Closing Remarks and Future Work

We have presented Uneek, some of its functions, and its potential to mitigate some of the methodological sufferings of linguistic labor. However, we see plenty of room for improvement. Here, we briefly mention two upcoming practical additions to Uneek: (i) *syntactic scope*, and (ii) a *multiple set analysis*.

(i) Sometimes, while faced with complex material, one would like to single out specific constituents of the parse tree for analysis, e.g. the subject. We plan to add functionality to Uneek to automatically extract these constituents. The user should be able to choose a constituent and get the annotation layers for its daughter nodes. This would considerably lessen the preprocessing of the input data.

(ii) We also plan to add a multiple set analys, enabling the user to get the intersection and difference between two or more sets. This would enable researchers and students to get results for complex comparative linguistic studies. Such an addition could come in handy soon, with the Multilingual FrameNet (MLFN) project underway. At the end of this project, Uneek could be used to answer some of the general MLFN questions below.[7]

1. ”Are some frames universal?”
2. ”Are there regular patterns of differences based on language families, regional groupings, etc.?”

The first question could then be answered by a multiple set intersectional analys of the annotation layers of language specific FrameNets. Uneek would automatically return their shared elements (frames, FEs, phrases, and so on). The second question may be answered by a multiple set uniqueness differentiation on sets of the FNs. Again, it would automatically return their unique elements.

Uneek is a simple tool, but sometimes there is strength in simplicity. Hopefully it will make the processing of complex data less tedious, enabling linguists to focus on the more interesting part of the field, i.e. coming up with explanations for linguistic phenomena.

5. 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.

6. Bibliographical References

Anthony, L. (2016). AntConc (version 3.4.4)[computer software]. Available at http://www. laurenceanthony.net Tokyo, Japan: Waseda University.

Borin, L., Forsberg, M., Hammarstedt, M., Rosén, D., Schäfer, R., and Schumacher, A. (2016). Sparv: Språkbanken's corpus annotation pipeline infrastructure. In *SLTC 2016. The Sixth Swedish Language Technology Conference, Umeå University, 17–18 November, 2016.*

Cantor, G. (1915). *Contributions to the Founding of the Theory of Transfinite Numbers*. Number 1. Open Court Publishing Company.

Das, D., Schneider, N., Chen, D., and Smith, N. A. (2010). SEMAFOR 1.0: A Probabilistic Frame-Semantic Parser. *Language Technologies Institute, School of Computer Science, Carnegie Mellon University*.

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

Fillmore, C. J. (1992). ”Corpus linguistics” vs. ”computer-aided armchair linguistics”. In *Directions in Corpus Linguistics: Proceedings from a 1991 Nobel Symposium on Corpus Linguistics*, Stockholm. Mouton de Gruyter.

Firth, J. R. (1962). *A Synopsis of Linguistic Theory, 1930-1955*. Basil Blackwell, Oxford, [1957] edition.

Harris, Z. S. (1954). Distributional Structure. *Word*, 10(2-3):146–162.

Manning, C. D., Surdeanu, M., Bauer, J., Finkel, J. R., Bethard, S., and McClosky, D. (2014). The Stanford CoreNLP Natural Language Processing Toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 55–60.

Nivre, J., de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R. T., Petrov, S., Pyysalo, S., Silveira, N., et al. (2016). Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*.

Ruppenhofer, J., Ellsworth, M., Petruck, M. R., Johnson, C. R., Baker, C. F., and Scheffczyk, J. (2016). FrameNet II: Extended Theory and Practice.

Scott, M. (2017). WordSmith Tools Help. http://www. lexically.net/downloads/version7/HTML/overview.html.

Torrent, T. T., da Silva Matos, E. E., Sigiliano, N. S., da Costa, A. D., and de Almeida, V. G. (2018). A flexible tool for an enriched FrameNet: the FrameNet Brasil Webtool. submitted for publication.

[7] https://framenet.icsi.berkeley.edu/fndrupal/node/5549

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.

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

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

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

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

[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) *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.

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

[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.

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]$_{\text{Descriptor}}$

Sources of Complexity in Semantic Frame Parsing for Information Extraction

Gabriel Marzinotto^{1,2}**, Frederic Bechet**^{2}**, Geraldine Damnati**^{1}**, Alexis Nasr**^{2}
(1) Orange Labs, (2)Aix Marseille Univ, CNRS, LIF
(1) Lannion France , (2) Marseille France
{gabriel.marzinotto, geraldine.damnati}@orange.com
{frederic.bechet, alexis.nasr}@lif.univ-mrs.fr

Abstract

This paper describes a Semantic Frame parsing System based on sequence labeling methods, precisely BiLSTM models with highway connections, for performing information extraction on a corpus of French encyclopedic history texts annotated according to the Berkeley FrameNet formalism. The approach proposed in this study relies on an integrated sequence labeling model which jointly optimizes frame identification and semantic role segmentation and identification. The purpose of this study is to analyze the task complexity, to highlight the factors that make Semantic Frame parsing a difficult task and to provide detailed evaluations of the performance on different types of frames and sentences.

Keywords: Frame Semantic Parsing, LSTM, Information Extraction

1. Introduction

Deep Neural Networks (DNN) with word embeddings have been successfully used for semantic frame parsing (Hermann et al., 2014). This model extends previous approaches (Das, 2014) where classifiers are trained in order to assign the best possible roles for each of the candidate spans of a syntactic dependency tree.

On the other hand, *recurrent neural networks* (RNN) with Long Short Memory (LSTM) cells have been applied to several semantic tagging tasks such as *slot filling* (Mesnil et al., 2015) or even frame parsing (Hakkani-Tür et al., 2016; Tafforeau et al., 2016) for Spoken Language Understanding. Currently, there is an important amount of research addressed to optimize architecture variants of the recurrent neural networks for the different semantic tasks. In SRL the current state of the art (He et al., 2017) uses an 8 layers bidirectional LSTM with highway connections (Srivastava et al., 2015), that learns directly from the word embedding representations and uses no explicit syntactic information.

More recently, (Yang and Mitchell, 2017) proposed a combined approach that learns a sequence tagger and a span classifier to perform semantic frame parsing making significant performance gains on the FrameNet test dataset.

However, there is little work in analyzing the sources of error in Semantic Frame parsing tasks. This is mainly due to the size of the SemEval07 corpus that contains 720 different frames and 754 Frame Elements (FE), with a lexicon of 3,197 triggers, for only 14,950 frame examples in the training set. Hence the size of the dataset, as well as number of examples per frame tends to be too small to perform this type of analysis. For this reason the analyses done by researchers in the domain focus mainly on the performance of their model on rare frames (Hermann et al., 2014). In (Marzinotto et al., 2018) a new corpus of French texts annotated following the FrameNet paradigm (Baker et al., 1998a) (Fillmore et al., 2004) is introduced. This new corpus has been partially annotated using a restricted number of Frames and triggers. The purpose was to obtain a larger amount of annotated occurrences per Frame with the counterpart of a smaller amount of Frames.

In this paper we focus on analyzing the factors that make a Frame hard to predict, describing which Frames are intrinsically difficult, but also which types of frame triggers are more likely to yield prediction errors and which sentences are complex to parse.

2. Sequence labeling model

2.1. Highway bi-LSTM approach

Following the previous work of (He et al., 2017), in this study, we propose a similar architecture, a 4 layer bidirectional LSTM with highway connections. For this model we use two types of LSTM layers, forward (F) layers and backward (B) layers which are concatenated and propagated towards the output using highway connections (Srivastava et al., 2015). A diagram of our model architecture is shown in Figure 1. There are 2 main differences between the model proposed in (He et al., 2017) and ours. First, we do not implement A* decoding of the output probabilities, second, our system not only relies on word embeddings as input features, but we also include embeddings encoding: syntactic dependencies, POS, morphological features, capitalization, prefixes and suffixes of the input words. We have observed these features to be useful for the FE detection and classification task.

In order to deal with both the multi-label and linking problems we have built training samples containing only one predicate. More precisely a sentence containing N predicates provides N training samples. The downside of this approach is that during prediction time, parsing a sentence with N predicates requires N model applications. At decoding time each pair { sentence , predicate } is processed by the network and a distribution probability on the frames and frame elements for each word is produced. To these probabilities we apply a *coherence filter* in which we take as ground truth the frame prediction (represented as the label assigned by the tagger to the trigger) and we discard frame element labels that are incompatible to the predicted frame.

3. The CALOR Semantic Frame Corpus

The experiments presented in this paper were carried out on the CALOR corpus, which is a compilation of doc-

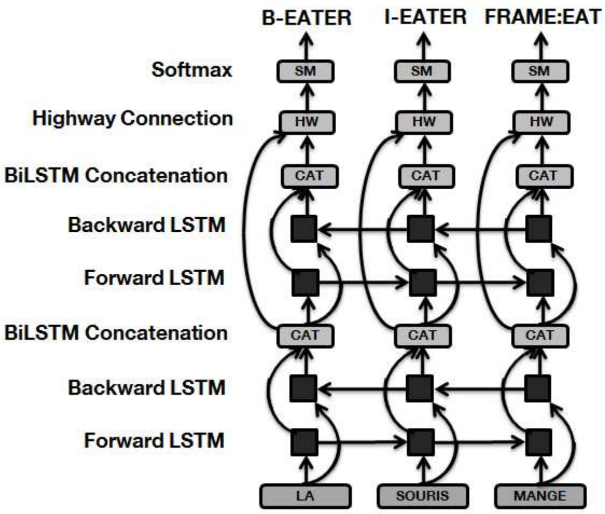


Figure 1: Highway bi-LSTM Model Diagram

uments in French that were hand annotated in frame semantics. This corpus contains documents from 4 different sources: Wikipedia's Archeology portal (WA, 201 documents), Wikipedia's World War 1 portal (WGM, 335 documents), Vikidia's [1] portals of Prehistory and Antiquity (VKH, 183 documents) and ClioTexte's [2] resources about World War one (CTGM, 16 documents). In contrast to full text parsing corpus, the frame semantic annotations of CALOR are limited to a small subset of frames from FrameNet (Baker et al., 1998b). The goal of this *partial parsing* process is to obtain, at a relatively low cost, a large corpus annotated with frames corresponding to a given applicative context. In our case this applicative context is Information Extraction (IE) from encyclopedic texts, mainly historical texts. Beyond Information Extraction, we attempt to propose new exploration paradigms through collections of documents, with the possibility to link documents not only through lexical similarity but also through similarity metrics based on *semantic frame structure*. The notion of document is more central in our study than in other available corpora. This is the reason why we have chosen to annotate a larger amount of documents on a smaller amount of Frames.

Precisely, while Framenet proposes 1,223 different frames, 13,635 LUs, and 28,207 frame occurrences on full text annotations, CALOR is limited to 53 different frames, 145 LUs (among which 13 are ambiguous and can trigger at least two frames) and 21,398 frame occurrences. This means that the average number of examples per frame in CALOR is significantly higher than in the full-text annotations from FrameNet.

4. Results

In order to run the experiments we divided the CALOR corpus into 80% for training and 20% for testing. This partition is done ensuring a similar frame distribution in training and test.

In the CALOR corpus, ambiguity is low, with only 53 different Frames. Most triggers have only 1 possible Frame. This makes the performance of our model in the frame selection subtask as high as 97%. For this reason we focus our analysis on the FE detection and classification subtask. We trained our model on the CALOR corpus and we evaluated it by thresholding the output probabilities in order to build the FE detection and classification precision recall curves shown in Figure 2. The three curves correspond to three possible precision-recall metrics: soft spans, weighted spans and hard spans. When evaluating using soft spans, a FE is considered correct when at least one token of its span is detected. In this case we achieve an F measure of 69,5%. If we use the weighted span metric, a FE is scored in proportion to the size of the overlap between the hypothesis and the reference segments. Using this metric we observe a 60,9% F-measure. Finally, the hard span metric considers a FE correct only if the full span is correctly detected. In this case, the performance degrades down to 51,7%. This experience shows that the model detects most of the FEs but it rarely finds the full spans. It should be possible to boost the model performance by +18pts of F-measure by expanding the detected spans to its correct boundaries.

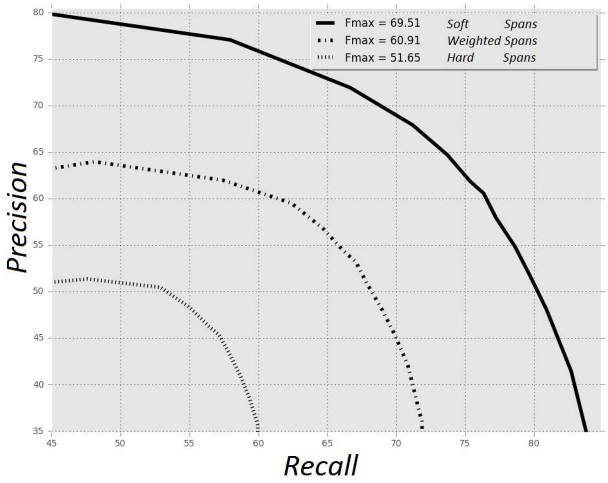


Figure 2: Model's Precision Recall curves using 3 different metrics: soft spans, weighted spans and hard spans

In the following subsections we present the results using the soft span metrics for the FE detection and classification task and we focus on analyzing several complexity factors in frame semantic parsing. We divide these factors into Frame Intrinsic (Section 4.1.) and Sentence Intrinsic 4.2.). In Section 4.3. we analyze the performance of the model at a document level using correlation analysis and regression techniques to retrieve relevant parameters that allow to predict the model performance on test documents.

4.1. Frame Intrinsic Complexity Factors

Some frames are intrinsically more difficult than others, this is due to their number of possible FEs, to the syntactic and lexical similarities between FEs and to the type of semantic concepts they represent. In Figure 3 we analyze the performance of our model on each FE with respect to their

[1] https://fr.vikidia.org

[2] https://clio-texte.clionautes.org/

Accomplishment	Activity-start	Age	Appointing	Arrest
Arriving	Assistance	Attack	Awareness	Becoming
Becoming-aware	Buildings	Change-of-leadership	Choosing	Colonization
Coming-to-believe	Coming-up-with	Conduct	Contacting	Creating
Death	Deciding	Departing	Dimension	Education-teaching
Existence	Expressing-publicly	Finish-competition	Giving	Hiding-objects
Hostile-encounter	Hunting	Inclusion	Ingestion	Installing
Killing	Leadership	Locating	Losing	Making-arrangements
Motion	Objective-influence	Origin	Participation	Request
Scrutiny	Seeking	Sending	Shoot-projectiles	Statement
Subjective-influence	Using	Verification		

Table 1: List of Semantic Frames annotated in the CALOR corpus

number of occurrences. In general, the more examples of a class we have, the better its performance should be. However, there are some ambiguity and complexity phenomena that must be taken into account.

4.1.1. Number of Frame Elements

The number of possible FEs is not the same for each Frame. Intuitively, a Frame with more FEs should be harder to parse. In table 2 we divide Frames into 3 categories *Small, Medium and Large* depending of their number of possible FEs. From this experience we observed that this is not such a relevant factor and that the number of possible FEs must be really large (above 10) in order to see some degradation in the model's performance.

	Nb Possible FEs	Fmeasure
Small Frames	1 to 7	70.5
Medium Frames	8 to 10	69.8
Large Frames	11 or more	65.8

Table 2: Performance for different Frame sizes

4.1.2. Specific Syntactic Realization

Some FEs have a specific syntactic realisation. Typically, the FEs that correspond to syntactic subjects or objects (also ARG0 or ARG1 in the PropBank Paradigm). This is the case of *Activity, Official, Sought Entity, Decision, Cognizer, Inspector, Theme, Hidden Object, Expressor and Projectile* , which show good performances even when the amount of training samples is reduced. On the other hand, there are FEs such as *Time, Place, Explanation, Purpose, Manner and Circumstances* , which are realized in syntax as modifiers and have a wider range of possible instantiations. For the latter the F-measure is much lower despite of a similar amount of training samples.

4.1.3. Syntax Semantic Mismatch

Some FEs syntactic realizations are different for different triggers. For example, the frame *Education Teaching* has `étudier` (*study*) , `enseigner`(*teach*) and `apprendre` (which can translate both into *learn* and *teach*) as potential triggers and *Student* , *Teacher* among their FEs. When the trigger is `étudier`, the syntactic subject is *Student*; when the trigger is `enseigner`, the subject is the *Teacher*; finally, when the trigger is `apprendre`, the subject could be either *Student* or *Teacher* and further disambiguation is needed in order to assign the correct FE. This explains the low performances observed in FEs such as *Teacher*.

Some FEs are very similar up to a small nuance. For example, the Frame *Education Teaching* has 6 possible FEs to describe what is being studied *Course, Subject, Skill, Fact, Precept and Role* and their slight difference relies in the type of content studied. This kind of FE are very prone to confusions even for human annotators. Moreover, if the FE is a pronoun, finding the correct label may not be possible without the sentence context.

Another type of FE similarity appears in symmetric actions, for example, the Frame *Hostile Encounter* , has FEs *Side1, Side2 or Sides* to describe the belligerents of an encounter. Such FEs are prone to confusions and for this reason, our model has a low performance on them.

4.2. Sentence Intrinsic Complexity Factors

We have identified three sentence intrinsic complexity factors, the Trigger POS, the Trigger Syntactic Position and the sentence length.

4.2.1. Trigger POS

As shown in table 3 the model performance varies more than 17pts of F measure depending on whether the triggers are nouns or verbs. This is due to the variety, in French, of the syntactic nature of Verb dependents when compared to Nouns. Verb arguments can be realized as Subjects, Objects, indirect Objects (introduced by specific prepositions), adverbs and Prepositional Phrases. Nouns arguments, in contrast, are usually realized as Prepositional Phrases and adjectives. Since FEs are mostly realized as arguments of their Frame trigger, Verb triggered Frames offer a wider range of syntactic, observable, means to distinguish its FE, making them easier to model.

4.2.2. Trigger Syntactic Position

We observe that the model's performance varies significantly depending on whether the sentence's triggers are at the root of the syntactic dependency tree or not. The results for this experience are summed up in Table 3. We observe that the easiest triggers are at the root of their syntactic tree and there is a difference of 14pts of F measure between them and the triggers that occupy other positions in the syntactic tree.

4.2.3. Sentence Length

Another important factor is sentence length. In general, longer sentences are harder to parse as they often present more FEs and a more complex structure. Also, it is in

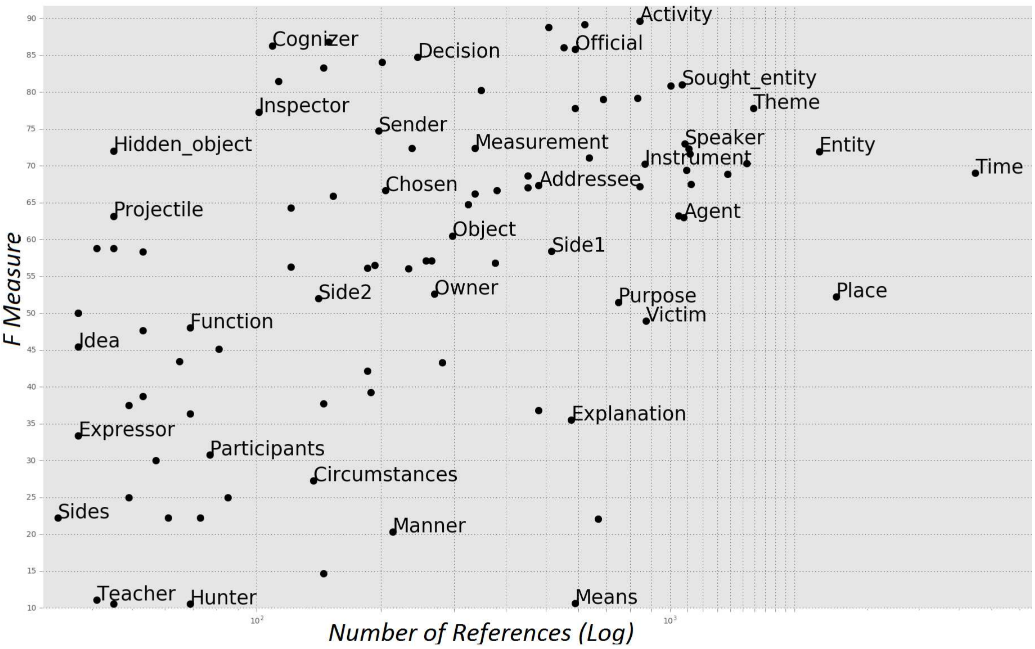


Figure 3: Model's performance for different FEs w.r.t the number of training samples

	Percentage	Fmeasure
Verbal Trigger	64.9%	75.2
Nominal Trigger	35.1%	57.5
Root Trigger	25.9%	79.5
Non Root Trigger	74.1%	65.4
All Triggers	100%	69.5

Table 3: Performance for different types of triggers

longer sentences that we find the most non-root triggers. In this experiment our model yields 74.0% F-measure for sentences with less than 27 words and 65.7% F-measure for sentences with 27 words or more. However, we also observed that parsing a long sentence (with 27 words or more) with the trigger at the root of the syntactic tree can still be done with a fairly good performance of 76.6%. While, when parsing a long sentence with non root triggers the performances degrade down to 62.8%. Table 4 presents the model performance with respect to trigger syntactic position and sentence length.

	Root Triggers	Non-Root Triggers
Short (< 27 words)	81.6	68.2
Long ($\geq$ 27 words)	76.5	62.8

Table 4: F-measure for different sentence lengths (Above and Below the Median) and trigger positions (Root and Non-Root)

All of these factors add up. Short sentences with a verbal trigger at the root of the syntactic tree can be parsed with an F-measure of 82.6% while long sentences with noun triggers that are not root of the syntactic tree are parsed with an F-measure of 52.8%.

4.3. Document Intrinsic Complexity Factors

In the previous sections we have presented the main factors that influence semantic frame parsing. In order to quantify the impact of each factor we have compared so far the performance of our model in subsets of the test set corresponding to the different modalities of the complexity factors. In this section, we try to directly evaluate the impact of these factors on the model performance. The analysis is performed here at the document level, the applicative motivation being to be able to predict the semantic parsing performances for a given new document.

We address the analysis of the complexity factors as a regression problem where we describe a dependent variable y (that quantifies the model performance) using a set of explanatory variables $X = (X_1, ..., X_n)$ which are our candidate complexity factors.

First, we use our model to generate hypothesis predictions of frame semantic parsing on the entire CALOR corpus, with a 5-fold protocole. For each fold, we train on 80% of the corpus, and generate predictions for the remaining 20%. Then, we evaluate the model's predictions using the gold annotations to compute the model's performance for each of the 735 documents in the CALOR corpus (Section 3.). For each of these documents, we also compute the set of features (complexity factors candidates) listed below:

- Percentage of root / non-root triggers.
- Percentage of verbal / nominal triggers.
- Mean phrase length.
- Mean trigger syntactic depth.
- Mean trigger position in sentence.

- Part of Speech (POS) distribution[1].
- Syntactic dependency relation (DEP) distribution[1].

To avoid taking into consideration parts of the document that were not processed by our semantic frame parser, these features are computed only using the sentences that contain at least 1 trigger. In order to make the analysis more robust to outliers we discarded the documents with less than 30 triggers, yielding a total amount of 327 documents. Unidimensional statistics show that the model's F-measure follows a Gaussian distribution across documents. This Gaussian distribution is centered at 69 pts of F-measure and has an standard deviation of 6.5 pts. This value of standard deviation shows that the model is fairly robust and has a stable performance across documents.

Finally, we used this set of document's features and model's performances in two experiments:

- To compute the Pearson correlation coefficient between the F-measure and each feature.
- To train a linear regression model that attempts to predict the model's performance on a document given a small set of parameters.

4.3.1. Pearson Correlation

We computed the Pearson correlation between each feature and the F-measure of the system and verified that the correlation coefficient passes the Student's t-Test. Table 5 shows the 15 parameters that have the highest absolute correlation with the F-measure.

	Rank	Pearson Correlation
Mean Trigger Depth	1	-0.44
Mean Trigger Position	2	-0.36
Verbal Trigger Percentage	3	$+0.31$
Mean Sentence Length	4	-0.30
DEP Oblique Nominal	5	$+0.30$
DEP Passive Auxiliary	6	$+0.29$
POS Punctuation	7	-0.28
POS Proper Noun	8	$+0.27$
POS Adverbs	9	-0.20
Multi Words Expressions	10	-0.20
DEP Prepositional Case	11	$+0.20$
POS Preposition	12	$+0.15$
POS Conjunction	13	-0.14
DEP Copula	14	-0.14
POS Number	15	$+0.11$

Table 5: Pearson Correlation between the best 15 Document Features and the F-measure

In table 5 we observe that the most important parameter is the syntactic depth of the trigger. As we have previously shown, triggers at the root of the syntactic tree have the best performances (also, root triggers are often verbs). The second most correlated parameter is the position of the trigger in the sentence. Triggers that are far from the beginning of the sentence show lower performance, as they are prone to errors in syntax. Our third and fourth parameters are the percentage of verbal triggers and the average sentence length. As shown in previous experiences, verbal triggers and short sentences are, in general, easier to parse.

This study also reveals morpho-syntactic parameters that are correlated with the model's performances: documents with a large amount of punctuation marks, adverbs, and conjunctions are more complex and harder to parse. On the other hand, documents with a large proportion of proper nouns are simpler, as proper nouns correspond to places, institutions and persons' names, which often appear as Frame Elements. The same observation can be made with Numbers, that correspond to dates and quantities. Prepositions also facilitate parsing, as they are associated to specific FEs.

As concerns dependency parsing related features, the highest correlation is observed for Oblique Nominal (OBL) dependency. OBL dependencies attach a noun phrase functioning as a non-core argument to the syntactic head. Documents with a large proportion of oblique nominal groups are positively correlated with the F-measure. OBL arguments are often annotated as FEs `(Time, Place, Purpose...)` and when they contain a Prepositional Case they are easy to associate to their corresponding FE. This also explains the positive correlation of the Prepositional Case. Surprisingly, documents with a large proportion of sentences in passive voice are correlated with better performances, while copula verbs degrade the results. In the CALOR corpus, some copula verbs are annotated as triggers *se nommer (to be named), être élu (to be elected), devenir (to become))* . Thus the negative correlation may be due to low performances for these lexical units.

Finally, Multi Words Expressions (MWE) are also associated with low performances, as the meaning of unseen MWE is harder to be inferred, misleading the semantic parser.

4.3.2. Performance Inference

In this experience we trained a linear regression model with incremental feature selection using cross validation. The objective is to predict the performances of our frame semantic parser on a document given a small set of parameters.

In incremental feature selection, we start with an empty set of selected features. At each iteration of the algorithm we test all the unselected feature candidates and we pick the feature that minimizes the cross validation mean square error (MSE) given all the previously selected features. The algorithm's stopping criterion finishes the process when the MSE no longer evolves. Unlike the previous experience where all features are evaluated independently, this experimet allows to select a smaller feature set that is not redundant.

We evaluate the usefulness of our linear regression models by comparing them with a naive constant model that always predicts the average document performance observed on the training corpus. Note that the training corpus here means the training corpus for regression estimation but not for the semantic frame parsing model estimation (each document is parsed in a k-fold protocole). Incremental Fea-

[1] POS and DEP from the Universal Dependencies project (http://universaldependencies.org/)

ture Selection determines that the optimal linear regression model is trained using 8 features and the insertion of more parameters does not reduce the MSE. Table 6 shows the MSE for the naive model `(Mean F-measure)` and compares it with each step of our linear regression with incremental feature selection. Each row in Table 6 adds a new feature to the linear regression model, up to the last row that contains the final set of 8 selected parameters.

We observe that the naive prediction algorithm has a MSE of 42.7. A linear regression model with only one feature `(Mean Trigger Depth)` reduces MSE by 16% relative and the best linear regression model with 8 features `(Mean Trigger Depth, DEP Oblique Nominal, Verbal Trigger Percentage, DEP Passive Auxiliary, DEP Copula, DEP Fixed Multi Words, POS Punctuation, POS Proper Noun)` yields a 41% relative MSE reduction.

Figure 4 shows a scatter plot of the documents with their predicted F-measure and their true F-measure. We can clearly observe that both scores are correlated and the variance that can be explained by the linear regression is $R^2 = 0.46$. However, there is still more than half of the variance that remains unexplained by the linear regression. This is because frame semantic parsing is a very complex task and the model's performances depend on many other phenomena such as the lexical coverage, lexical units and frames that appear within a document, the type of FEs that are evocated and the degree of ambiguity at each level.

	# Features	MSE
Mean F-measure	0	42.7
Mean Trigger Depth	1	35.9
DEP Oblique Nominal	2	33.5
Verbal Trigger Percentage	3	30.5
DEP Passive Auxiliary	4	29.1
DEP Copula	5	27.4
Multi Words Expressions	6	26.3
POS Punctuation	7	25.6
POS Proper Noun	8	25.1

Table 6: Mean Squared Error (MSE) for Linear Regression with Incremental Feature Selection

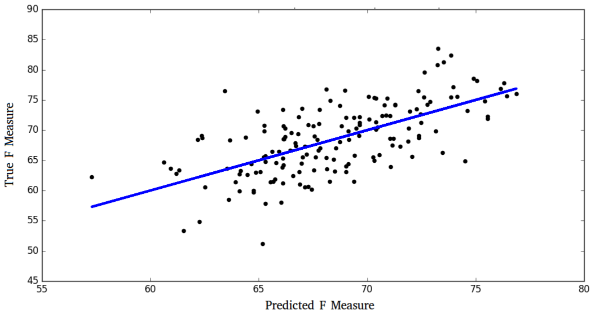


Figure 4: Document Scatter plot showing True F-measure vs Predicted F-measure using a Linear Regression with 8 features

5. Conclusion

In this paper we proposed to identify complexity factors in Semantic Frame parsing. To do so we ran experiments on the CALOR corpus using a frame parsing model that considers the task as a sequence labeling task. In our case only *partial* annotation is considered. Only a small subset of the FrameNet lexicon is used, however the amount of data annotated for each frame is larger than in any other corpora, allowing to make more detailed evaluations of the error sources on the FE detection and classification task. The main contribution of this work is to characterize the principal sources of error in semantic frame parsing. We divide these sources of error into two main categories: Frame intrinsic and sentence intrinsic. Examples of Frame intrinsic factors are the number of possible FE, and the syntactical similarity between them. As for the sentence intrinsic factors, we enhanced the position of the trigger in the syntactic tree, the POS of the trigger and the sentence length. In this work we showed that some morpho-syntactic categories and syntactic relations have an impact on the complexity of the frame semantic parsing. Finally, we showed that it is possible to make a fair prediction of the model's performance on a given document thanks to a regression estimation knowing its sentence intrinsic parameters. The features selected for the regression estimation confirm the observations regarding the task complexity but also enhance new assertions. The complexity factors presented in this article may allow further work on feature engineering to improve the frame semantic parsing models.

6. Bibliographical References

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998a). The berkeley framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics-Volume 1*, pages 86–90. Association for Computational Linguistics.

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998b). The berkeley framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, ACL '98, pages 86–90, Stroudsburg, PA, USA. Association for Computational Linguistics.

Das, D. (2014). Statistical models for frame-semantic parsing. In *Proceedings of Frame Semantics in NLP: A Workshop in Honor of Chuck Fillmore*, volume 1929, pages 26–29.

Fillmore, C. J., Baker, C. F., and Sato, H. (2004). Framenet as a "net". In *LREC*.

Hakkani-Tür, D., Tur, G., Celikyilmaz, A., Chen, Y.-N., Gao, J., Deng, L., and Wang, Y.-Y. (2016). Multi-domain joint semantic frame parsing using bi-directional rnn-lstm. In *Proceedings of The 17th Annual Meeting of the International Speech Communication Association.*

He, L., Lee, K., Lewis, M., and Zettlemoyer, L. (2017). Deep semantic role labeling: What works and what's next. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics.*

Hermann, K. M., Das, D., Weston, J., and Ganchev, K. (2014). Semantic frame identification with distributed word representations. In *ACL (1)*, pages 1448–1458.

Marzinotto, G., Auguste, J., Bechet, F., Damnati, G., and Nasr, A. (2018). Semantic frame parsing for information extraction : the calor corpus. In *LREC*.

Mesnil, G., Dauphin, Y., Yao, K., Bengio, Y., Deng, L., Hakkani-Tur, D., He, X., Heck, L., Tur, G., Yu, D., et al. (2015). Using recurrent neural networks for slot filling in spoken language understanding. *IEEE/ACM Transactions on Audio, Speech and Language Processing (TASLP)*, 23(3):530–539.

Srivastava, R. K., Greff, K., and Schmidhuber, J. (2015). Highway networks. *CoRR*, abs/1505.00387.

Tafforeau, J., Bechet, F., Artiere, T., and Favre, B. (2016). Joint syntactic and semantic analysis with a multitask deep learning framework for spoken language understanding. *Interspeech 2016*, pages 3260–3264.

Yang, B. and Mitchell, T. (2017). A joint sequential and relational model for frame-semantic parsing. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1247–1256. Association for Computational Linguistics.

The Danish FrameNet Lexicon: Method and Lexical Coverage

Sanni Nimb
The Society for Danish Language and Literature
sn@dsl.dk

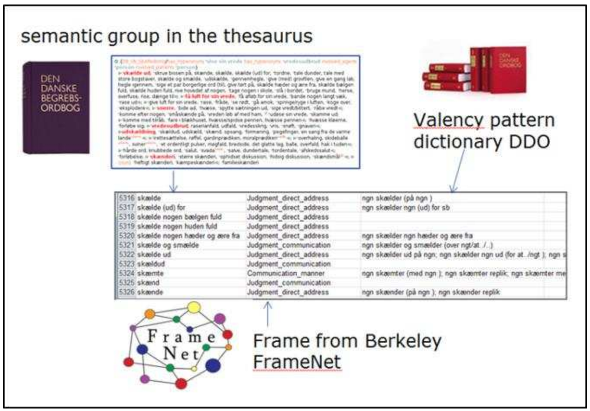
Abstract

This paper presents and discusses the results of compiling a comprehensive Danish frame lexicon compliant with the Berkeley FrameNet standard by making use of linked lexical data from two Danish resources, namely the semantic and thematic grouping of verbs and verbal nouns in a Danish thesaurus with the valency patterns of the same verbs in a monolingual Danish dictionary. The frame lexicon covers a large number of Danish lemmas, including phrasal verbs and multiword units, and furthermore gives information on one or more phrases illustrating the typical textual context in which the lemma evokes the frame in question. The overall aim is to supply annotators of semantic frames and roles in Danish texts in future research projects with a restricted and thereby manageable set of possible frames to choose from. We present the content of the lexicon in detail, including a comparison with the frame coverage of Berkeley FrameNet.

Keywords: thesaurus, frame lexicon, Danish

1. Lexical resources as input

In order to compile a Danish frame lexicon compliant with the international standard resource Berkeley FrameNet (Ruppenhofer et al., 2016, henceforth BFN) we combine the valency information on verbs in a comprehensive monolingual dictionary with the semantic and thematic grouping of the same verbs and related verbal nouns in a thesaurus. The dictionary we use (*Den Danske Ordbog*, henceforth the DDO dictionary[1]) contains approx. 100,000 lemmas and 136,000 senses. The thesaurus (*Den Danske Begrebsordbog* ('The Danish Concept Dictionary', Nimb et al., 2014 a & b, henceforth the thesaurus) is based on and linked to the lemma senses in the DDO dictionary and covers 80 % of the senses. The links between the two resources allow us to combine all sorts of lexical information and use it for different purposes, in this case semantic relatedness on the one hand and syntactic information on the other hand used as input to the manual assignment of frame information, see figure 1. To a high degree, the valency patterns in the DDO dictionary reflect the semantic role inventory as described in BFN, and thereby help us select and assign the most appropriate frame.


semantic group in the thesaurus
DEN DANSKE BEGREBS-ORDBOG
Valency pattern dictionary DDO
Frame Net
Frame from Berkeley FrameNet

5316	skælde	Judgment_direct_address	ngn skælder (på ngn)
5317	skælde (ud) for	Judgment_direct_address	ngn skælder ngn (ud) for sb
5318	skælde nogen bælgen fuld	Judgment_direct_address	
5319	skælde nogen huden fuld	Judgment_direct_address	
5320	skælde nogen hæder og ære fra	Judgment_direct_address	ngn skælder ngn hæder og ære fra
5321	skælde og smælde	Judgment_communication	ngn skælder og smælder (over ngt/at ../..)
5322	skælde ud	Judgment_direct_address	ngn skælder ud på ngn; ngn skælder ngn ud (for at../ngt); ngn s
5323	skældud	Judgment_communication	
5324	skæmte	Communication_manner	ngn skæmter (med ngn); ngn skæmter replik; ngn skæmter me
5325	skænd	Judgment_communication	
5326	skænde	Judgment_direct_address	ngn skænder (på ngn); ngn skænder replik

Figure 1: Linked data: The word groups in a Danish thesaurus combined with the valency information in a Danish dictionary constitute the background for the framenet

The semantically related verbs and verbal nouns in the thesaurus are typically assigned one of a rather restricted set of BFN frames, making it possible instantaneously to compile large amounts of lemmas and expressions within the same semantic area. Due to the close relation between English and Danish we assume that the frame descriptions and the role inventory from BFN can be transferred directly and used in future Danish annotation tasks. This was confirmed when testing part of the frame lexicon in a pilot project: we did not encounter any problematic cases of role assignment (Nimb et al., 2017; Pedersen et al., 2018). The frame lexicon project was carried out at DSL in collaboration with the University of Copenhagen and funded by the Carlsberg Foundation 2016-2017.

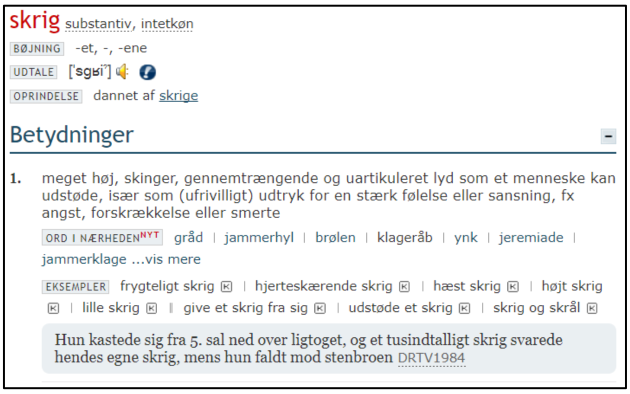
2. The method

We use the source xml-structured document of the thesaurus which arranges the Danish vocabulary in 22 chapters and 888 named sections[2] and furthermore in an average of 9-10 annotated semantic groups in each section where words and expressions are grouped according to semantics, not word classes (Nimb et al., 2014 b), making it very useful for our purpose since verbs and their corresponding verbal nouns are grouped together. Since all semantic groups are formally annotated with coarse-grained semantic information (i.e. 'property', 'act', 'event', 'person' etc.), we can easily identify and thereby focus on acts and events exclusively in order to create a core frame lexicon describing the part of the Danish vocabulary that occurs with semantic roles. If we take a closer look at the different semantic groups of a section, the section with the title 'Crying', for example, contains one group with words meaning 'person who cries', formally annotated with the type 'person'. Another group represents the verbs and verbal nouns with the meaning 'to cry', and is annotated with 'act/involved agent' information, while a third group lists adjectives describing persons who (easily) cry. In total there are approx. 8,300 semantic groups in the thesaurus. 1/5 of these groups (1487 semantic groups), containing more than 42,000 words and expressions, were identified as being of the type 'act' or 'event' via the coarse-grained formal semantic annotations of each group.

[1] DDO was compiled as a printed dictionary in the 1990s. Today the dictionary is online and regularly extended with new words and expressions.

[2] The section and chapter division is inspired by a German thesaurus (Dornseiff, 2004), but adjusted to the Danish language community of today.

The vocabulary of the groups includes not only single lemmas but also many of the collocations, for example support verb constructions, which are described in the DDO dictionary based on corpus statistics. In the case of the verbal noun *skrig* (‘scream’), sense 1 in DDO (see figure 2), we find the collocational expressions *give et skrig fra sig* (‘give/utter a cry’) and *udstøde et skrig* (‘utter a cry/cry out/shriek’), and they are both included in the thesaurus data and thereby also assigned their corresponding frames from BFN when compiling the frame lexicon.


skrig substantiv, intetkøn

BØJNING -et, -, -ene

UDTALE [ˈsgʁiˀ]

OPRINDELSE dannet af skrige

Betydninger

1. meget høj, skinger, gennemtrængende og uartikuleret lyd som et menneske kan udstøde, især som (ufrivilligt) udtryk for en stærk følelse eller sansning, fx angst, forskrækkelse eller smerte

ORD I NÆRHEDEN NYT gråd | jammerhyl | brølen | klageråb | ynk | jeremiade | jammerklage ...vis mere

EKSEMPLER frygteligt skrig | hjerteskærende skrig | hæst skrig | højt skrig | lille skrig ‖ give et skrig fra sig | udstøde et skrig | skrig og skrål

Hun kastede sig fra 5. sal ned over ligtoget, og et tusindtalligt skrig svarede hendes egne skrig, mens hun faldt mod stenbroen DRTV1984

Figure 2: The entry of the verbal noun *skrig* (‘scream’) in the DDO dictionary with several collocations (listed after EKSEMPLER (‘examples’)). Apart from the lemma itself, two of these are included in the same semantic group in the thesaurus: *give et skrig fra sig* (lit. ‘give a scream from oneself’) and *udstøde et skrig* (‘give a scream’), both meaning ‘to scream’.

Many words and expressions in the DDO dictionary are part of more than one section in the thesaurus and therefore also listed more than once in the extracted data material used as input to the frame assignment. E.g. the verb *guide* (‘to guide’) is part of 4 different sections in 3 different chapters in the thesaurus and has therefore been assigned four frame values in the frame lexicon: Assistance, Cotheme, Leadership and Telling. Likewise the verb *cruise* (‘to cruise/move easily’) which occurs in seven different sections (in three different chapters) has been assigned five different frames, namely Motion, Self_motion, Operate_vehicle, Finish_competion (in the sense ‘to win easily in sports’) and Personal_relationship (in the sense ‘to search for a partner’). The fact that the collocations in the DDO dictionary are statistically corpus-based and that representation in more than one thesaurus section often reflects different aspects of the same sense, similar to what a selection of corpus examples would do, allows us to consider the extracted data as a sort of ‘condensed’ corpus data in the form of small representative bits of phrases we would typically find in Danish texts if we were going to annotate a set of corpus examples with frames from BFN.

The thesaurus data and the corresponding valency patterns from the DDO dictionary were identified and extracted into a spreadsheet (carried out by Thomas Troelsgård, DSL). In figure 3 we present a small extract of the combined data, including the frames that we manually assigned after having translated them into English by use of a Danish English dictionary and afterwards having looked up the lexical_unit equivalents and their corresponding frames in BFN.

From the thesaurus : word/expression with the meaning ‘to cry/to scream’	Shared sense id number	From the DDO dictionary: valency pattern ‘somebody cries (+ manner) (because of something)’	Assigned frame from BFN
klage sig	21034458	ngn klager (sig) over ngt	Make_noise
jamre	21074699	ngn jamrer (sig) (over ngt)	Make_noise
jamre over	21074699	ngn jamrer (sig) (over ngt)	Judgment_communication
jamre sig over	21090433	ngn jamrer (sig) (over ngt)	Judgment_communication
græde	21074701	ngn græder (+ måde) (af noget)	Make_noise
skrige	21074700	ngn skriger (+ måde) (af noget)	Make_noise
skrige af smerte	21074700	ngn skriger (+ måde) (af noget)	Make_noise
give et skrig fra sig	21074701	NONE (*skrig* = verbal noun)	Make_noise
udstøde et skrig	21010806	NONE (*skrig* = verbal noun	Make_noise
sætte i et hyl	21033375	NONE (*hyl* = verbal noun)	Make_noise

Figure 3: The spreadsheet with the extracted data, in this case words and expressions with the meaning ‘to cry’ (some are verbs, others support verb constructions), linked to their corresponding valency patterns from the DDO dictionary via shared id numbers. The right colon presents the assigned frames from BFN.

In an initial pilot project (Nimb et al. 2017) we focused on the vocabulary from only two semantic areas, namely communication and cognition. The sections and semantic groups covering these two areas were easy to identify in the thesaurus due to the chapter names, and constitute approx. 16 % of all act and event groups in the thesaurus. We assigned a total of 104 different frames and tested the data in an annotation task where the supersenses of the verbs (verb.communication or verb.cognition) were already manually identified. The overall conclusion was that the compilation method was very efficient. By focusing on one semantic area at a time, which was made possible via the section and chapter grouping in the thesaurus, the lexical data considered was likely to be assigned the same frame, or at least a closely related frame, from BFN. When annotating with the frames, the decision-making was largely facilitated by the restricted number of possible frames for each verb in the text. But we also concluded that some of the most frequent verbs were lacking important frame values due to the fact that not all senses of highly polysemous verbs were represented in the thesaurus. Nimb et al. (2017) describes the pilot project in detail. In this paper, however, we focus on the lexical coverage of the entire Danish frame lexicon.

3. Lexical coverage and the distribution of frames

The lexicon consists of 23,260 unique words or expressions. Of these, 12,142 are single lemmas (e.g. the verb *chartre* ('to chart') and the noun *chartring* ('charting')), while 11,118 are expressions from the DDO dictionary consisting of two or more words. Most of these are fixed verbal expressions, including phrasal verbs, but we also find lexical collocations, mostly verbs with a typical object (*nippe til maden* ('to pick at the food'), *mene det modsatte* ('to mean the opposite')) or nouns with a typical support verb (*fatte en beslutning* ('to make a decision')). We also find verb phrases with one or more obligatory arguments represented by pronouns (e.g. *tale noget igennem* ('to talk something through'). The single lemmas and the fixed expressions correspond to Lexical Units in BFN (e.g. 'give up' and 'nip in the bud' are fixed expressions in BFN). We estimate that the frame lexicon covers approx. 20,000 Lexical Units, but it should be mentioned that the borderline between fixed expressions and collocations is elusive.

There are 33,930 unique combinations of word/multiword expression and frame value. A lemma or expression might be included two times or more in the lexicon, depending on how often it is represented in the different thesaurus sections. Due to this, there are 42,270 combinations of word/expression + frame value + thesaurus group number. The group numbers represent a different and often more fine-grained semantic relatedness of the data than the frame divisions do and are e.g. useful in the case of negative/positive words within the same frame group. I.e. they make it possible to divide words with the frame Remembering_experience into two groups, those meaning 'to forget' and those meaning 'to remember'.

Lemma in DDO dict.	Sense in DDO dict.	Sense also SynSet member in the Danish WordNet DanNet	Unique lemmas + expressions with frame value	Lemmas + expressions with unique combination of frame value and thesaurus group number
12,142	21,812	6,877	33,930	42,270

Table 1: Statistics on data in the frame lexicon

	Lemmas from DDO dict.	Senses from DDO dict.	Also in the Danish WordNet DanNet	Frame values
nouns	6,490	8,372	2,063	11,032
verbs	5,300	12,354	4,750	17,731

Table 2: Statistics on nouns and verbs in the frame lexicon

The words and expressions which are represented stem from 20,820 different senses from 12,124 lemmas. Some of the covered senses are also linked to synsets in the Danish WordNet DanNet (Pedersen et al., 2009), namely 38 % of the 12,354 verb senses and 25 % of the 8,372 noun senses. The 5,300 different verb lemmas have altogether been assigned 17,731 frame values. This means that 80 % of the verb lemmas in the DDO dictionary are represented in the frame lexicon with an average of 3.3 frames per verb. If we look at nouns, a total of 6,490 are represented in the lexicon and assigned a total of 11,032 frames (1.7 frames per noun). In tables 1 and 2 we present some statistics, and in figure 4 we list a small part of the frame lexicon entries, exemplified with a selection of Danish verbs originating in the English language.

	Lemma		Frame
v.	*chartre*	'to rent a plane or boat'	Renting
n.	*chartring*	'renting a plane or boat'	Renting
v.	*chatte*	'to chat via internet'	Communication_ means
v.	*chippe*	'to move a ball'	Cause_motion Sports_jargon
v.	*coache*	'to guide wrt personal career'	Education_teaching
v.	*crashe*	'to have an accident by car'/ 'to hit' / 'to participate in a party without being invited'	Catastrophe Impact Drop_in_on

Figure 4: A small extract from the Danish frame lexicon sorted by lemma. The frames are assigned to both verbs and verbal nouns. Valency patterns are not included in the release of the lexicon, but example phrases and group numbers from the thesaurus are (not illustrated here).

Many, but not all the nouns are both single lemmas in the lexicon (with one or more frame assignments) as well as part of a verbal phrase, typically combined with a support verb (with one or more frame assignments). Also a number of adjectives and adverbs are represented in the lexicon but in this case always as part of a verbal phrase. In both cases the verb in the verbal phrase is identified in a specific data field.

3.1 The frame inventory used for Danish

671 different frame values from BFN (~2/3 of all frame values) have been assigned to the Danish vocabulary. We have not yet compared the two sets of frames but plan to do so in order to identify English frames which have not been applied. We expect to find cases where such frames might have been better choices. Due to our method the lexicographer became rather confident with the different frame possibilities in BFN, and this guarantees at least to a certain degree that the Danish frame assignment is homogeneous across the lexicon.

The most frequent frame in the Danish lexicon is Self_motion (2% of the data). Subsequently, we find Experiencer_focused_emotion, Statement, Stimulate_ emotion, Judgment_communication, and Cogitation (all between 1 and 2% of the data). An additional 36 frames are assigned to between 0,5 and 1 % of the data, covering areas such as sports (Sports_jargon), acts in general (Removing, Filling, Processing_materials, Bungling, Intentionally_act), eating and drinking (Ingestion), and communication (Text_creation, Request, Respond_to_proposal). The remaining frames (~ 630) used to describe the Danish verbs and verbal nouns are only applied on less than 0.5 percent of the vocabulary in the thesaurus, respectively. Of these, almost 200 are used only 10 or less times, and approx. 50 are used only once.

At a first glance into the statistics of the applied frames for Danish, the most frequently used ones describe the semantic areas of motion, emotion, act, communication and cognition. In supersense-annotated Danish texts (Martínez et al. 2015) act, communication and cognition are also among the most frequent, while motion and emotion are less frequent. While the far most frequent verb sense in texts is 'stative', we do not find a large variety of lemmas or frames with this sense in neither the thesaurus, nor the frame lexicon. Similarly, there are only a few lemmas and frames with the sense 'possession' in lexicons compared to the high frequency of the sense in corpora.

3.2 Semantic areas covered by the Danish thesaurus but not (yet) by Berkeley FrameNet

Due to the fact that the Danish thesaurus represents more or less the entire vocabulary of a comprehensive corpus-based Danish dictionary which covers all general semantic areas in Danish, it is interesting to compare its coverage with BFN and study the cases where it was difficult to find corresponding frames to assign to the Danish words. In figure 5 we list the cases where we found it hard to find appropriate frames for one or maybe more verbs with a given sense in Danish, either because English conceptualization seems to differ from Danish, or because BFN does not cover the sense yet. It should be mentioned that we still need to study the cases in more detail and validate the data in order to find out whether we have simply misinterpreted the coverage of already existing frames in BFN. We exclude cases in which BFN states that frames are planned to be created (e.g. acts in sports and many scientific domains).

Areas and concepts covered by the Danish thesaurus, but not (yet) by Berkeley FrameNet
Not a human act
a calm situation (note: opposite to the frame Chaos); to go well, to be solved (note: about situation/problem); a machine carrying out a function; biological reproduction (note: both animals and plants); plants growing; animals living and acting
'General' human acts
to have a habit/to carry out a habit; to delimit something; to exaggerate when carrying out an activity, to overdo something; to hurry when carrying out an activity; to repeat an activity
Cogitation
to change your opinion; to mentally accept/adapt to something
Cleaning/polluting/recycling
to make something clean (note: the frame Removing is too broad in its sense, we find); to ventilate/clean out the air; to make something dirty, to pollute; to throw out something (note: as garbage); to protect nature (note: we have used Protecting but find it too broad); to recycle something/reuse
Creation
knitting, sewing etc.; concrete repairing (note: in both cases we find Processing_ materials too broad)
Social acts
to force somebody to do something (note: without using violence), to defend somebody (note: by speaking, not physically); to mediate/act as a mediator; to celebrate something; to meet somebody by coincidence; to stay in a place without staying overnight; to feed/give food to other persons; to take care of children/to babysit; the act of flirting with somebody
Body activities
to do sports, run, ride, surf (note: without competing, focus instead on pleasure/health purposes); to play games, to play for fun; gambling; to bathe for fun, e.g. in the sea; the act of masturbating; to go to bed (note: to get up is covered); not to eat/to be on a diet; to do nothing, to relax
Domain-specific acts
to plant trees, flowers, foresting (note: the frame Agriculture is too narrow, we find); sterilization of animals; to dig, to make holes; to parcel out/subdivide a piece of land; economics : raise money on; mortgage; laws : defend in the court
Supernatural acts/events
to practice witchcraft, to conjure; to haunt a place; to tell fortunes

Figure 5: Cases where we found it hard to find appropriate frames in BFN

4. Conclusions and future work

The freely available lexicon which can be downloaded at https://github.com/dsldk/dansk-frame-net contains data on the lemma, its word class and its frame value, a typical phrase or collocation, and the group number from the thesaurus. In cases of noun lemmas with verbal phrase examples, the verb is furthermore identified. The valency patterns from the DDO dictionary are not part of the data.

While the number of different verb lemmas is very high in the thesaurus and thereby also in the frame lexicon, verb polysemy as it is represented in the DDO dictionary is less extensively covered. We therefore plan to supply especially the highly polysemous verbs - which are also the ones occurring very often in texts - with more frames, and in this case base the compilation on the DDO dictionary's sense descriptions. So far only the frames regarding cognition and communication have been used for annotation. Our hope is to use the frame lexicon in future annotation projects. We furthermore plan to integrate the frame data in the Danish WordNet (which is also linked to the senses of the DDO dictionary) and use the frame values to improve the hierarchies of verbs in the WordNet.

5. Bibliographical References

The Danish Dictionary (DDO dictionary) (2008-): online dictionary, ordnet.dk/ddo, Det Danske Sprog- og Litteraturselskab, Copenhagen, Denmark.

DanNet: andreord.dk; wordnet.dk.

Danish English Dictionary: ordbog.gyldendal.dk, Gyldendal, Copenhagen.

Dornseiff, Franz (2004) *Der deutsche Wortschatz nach Sachgruppen*, W. De Gruyter, Berlin; New York.

Martínez Alonso, H., Johannsen, A., Olsen S., Nimb S., Sørensen N., Braasch, A., Søgaard, A. & Pedersen, B. S. (2015). Supersense tagging for Danish. In: *Proceedings of the 20th Nordic Conference of Computational Linguistics NODALIDA 2015. Vol. 109*, Linköping University Electronic Press, NEALT Proceedings Series, Vol. 23.

Nimb, S., Lorentzen, H., Theilgaard, L., Troelsgård, T. (2014 a). *Den Danske Begrebsordbog*, Det Danske Sprog- og Litteraturselskab, Copenhagen, Denmark.

Nimb, S., Lorentzen H., Trap-Jensen, L. (2014 b). The Danish Thesaurus: Problems and Perspectives. In: Abel A., Vettori C. & Ralli N. (eds.). *Proceedings of the XVI EURALEX International Congress: The User in Focus*. 15-19 July 2014. Bolzano/Bozen 2014: EURAC Research, pp. 191-199.

Nimb, S., A .Braasch, S.Olsen, B. S. Pedersen, A. Søgaard (2017). From thesaurus to framenet. In: *Proceedings of eLex 2017*, Leiden.

Pedersen, Bolette Sandford; Nimb, Sanni; Asmussen, Jørg; Sørensen, Nicolai; Trap-Jensen, Lars; Lorentzen, Henrik (2009) DanNet - the challenge of compiling a WordNet for Danish by reusing a monolingual dictionary. In: *Language Resources and Evaluation, Vol. 43*, p. 269-299.

Pedersen, B.S., Nimb, S., Søøgaard, A., Hartmann, M., Olsen, S. (2018) A Danish FrameNet Lexicon and an annotated Corpus used for Training and Evaluating a Semantic Frame Classifier. In: *Proceedings of LREC 2018*.

Ruppenhofer, J., Ellsworth, M., Petruck, M. R. L., Johnson, C. R., Baker, C. F., Scheffczyk, J. (2016). *FrameNet II: Extended Theory and Practice* (Revised November 1, 2016.) https://framenet.icsi.berkeley.edu/fndrupal/the_book.

Linking Japanese FrameNet with Kyoto University Case Frames Using Crowdsourcing

Kyoko Hirose Ohara, Daisuke Kawahara, Satoshi Sekine, Kentaro Inui
Keio University/RIKEN, Kyoto University/RIKEN, RIKEN, Tohoku University/RIKEN
ohara@hc.st.keio.ac.jp, dk@i.kyoto-u.ac.jp, satoshi.sekine@riken.jp, inui@ecei.tohoku.ac.jp

Abstract

We report on an ongoing project to link Japanese FrameNet (JFN) annotated sentences and Kyoto University Case Frames (KCF) example sentences that share the same meaning of a Japanese predicate (i.e., a verb, an adjective, or an adjectival noun), by way of crowdsourcing. JFN assigns a "cognitive frame" (a script-like conceptual structure that describes a particular type of situation, object, or event along with its participants and props, assumed in the theory of Frame Semantics) to each sense of Japanese words (mostly verbs, adjectives, adjectival nouns, and nouns). On the other hand, each "case frame" in KCF is a predicate-argument structure. Whereas JFN has been constructed manually so far, KCF was automatically constructed from 10 billion Japanese sentences taken from Web pages. By linking JFN annotated sentences and KCF example sentences that share the same meaning of a predicate, we can ultimately increase the size of JFN and also add semantic information to KCF. We use JFN cognitive frames to link the sentences in the two resources. We crowdsourced this task to ensure rapid and large-scale mappings between the two. Our preliminary results suggest that the proposed crowdsourcing method for linking the resources via cognitive frames is promising.

Keywords: Kyoto University Case Frames, Japanese FrameNet, Crowdsourcing

1. Introduction

We report on a project to link Japanese FrameNet (JFN) annotated sentences and example sentences in Kyoto University Case Frames (KCF) that share the same meaning of a Japanese predicate (i.e., a verb, an adjective, or an adjectival noun), by way of crowdsourcing.

There are two types of so-called frame knowledge. The first type concerns dividing what speakers know about the world into "cognitive frames," that is, script-like conceptual structures that describe a particular type of situation, object, or event along with its participants and props, in a top-down manner. The second type of frame knowledge involves predicate-argument structures and describes, in a bottom-up fashion, what kinds of arguments individual predicates (mostly verbs, including copulas, and adjectives) take, i.e., "case frames."

Both kinds of frame knowledge, that is, cognitive frames ("top-down frame knowledge") and case frames ("bottom-up frame knowledge"), have been organized into language resources and have become fundamental to text understanding. An example of the former is FrameNet (FN), an English language resource that relates cognitive frames to individual English words (mostly verbs, nouns, and adjectives) (Fillmore and Baker, 2010; Ruppenhofer et al., 2016). FN also includes corpora annotated with information about cognitive frames that words evoke. Resources similar to FN have also been built for languages other than English by manual elaboration or translation. However, they often have a problem in coverage, since most of them use a partial set of the cognitive frames defined in FN. For example, JFN, which has been constructed manually, has a smaller set of cognitive frames and smaller numbers of Lexical Units (LUs) and annotated sentences than the original FN, as shown in Table 1.

KCF is an example of the latter type of language resources, which has been automatically acquired from a large raw corpus of Japanese (Kawahara et al., 2014). It has a wide coverage and statistical information. However, although KCF applies a clustering algorithm to generate case frames with different usages, it does not contain semantic information.

	FN	JFN
# of Cognitive Frames	1223	979
# of Lexical Units (LUs)	13638	5029
# of Annotated Sentences	202229	7899

Table 1: Comparison of FrameNet (FN) and Japanese FrameNet (JFN)

This paper proposes a method to link JFN and KCF to exploit the advantages of both resources. There have been no attempts to combine a resource containing top-down frame knowledge (i.e., cognitive frames) with bottom-up frame knowledge (i.e., case frames). By using our method, it is possible to build a wide-coverage knowledge resource of cognitive frames using statistical information.

Our method links an automatically acquired case frame in KCF with one of the JFN cognitive frames associated with each verb, adjective, or adjectival noun (hereafter "predicate") in Japanese. To conduct this task fast and on a large scale, we employ the crowdsourcing technique. Specifically, for each predicate, we ask crowdworkers to link an example sentence of a KCF case frame with an example sentence of a JFN cognitive frame. One reason for using example sentences is to facilitate the linking task for crowdworkers. Another is to enable the reuse of the linking knowledge for newly reconstructed case frames. In fact, KCF case frames are often reconstructed by improving the clustering algorithm and by expanding the size of a source corpus.

Our method seems to be promising in the following three aspects:

- To scale up the size of sentences annotated with cognitive frames in JFN;
- To facilitate identifying missing cognitive frames in JFN;

- To add new LUs to existing cognitive frames in JFN.

Our ultimate goals include: increasing the size of JFN; and adding semantic information to each case frame in KCF. Our first step, however, whose preliminary results are reported in this paper, involves matching JFN annotated sentences and KCF example sentences that share the same JFN cognitive frame, in other words, assigning a JFN cognitive frame to each KCF case frame.

The inherent difficulty and complexity of the FN annotation processes have prompted researchers in the FN community to look for ways to expand the database of annotated sentences. One idea is to reuse some of the work done by other projects. There are, however, few language resources that share some of the principles of Frame Semantics in general and of FrameNet in particular.[1] We will argue, however, that linking JFN and KCF is indeed possible, since KCF does not include semantic information incompatible with the principles of Frame Semantics and since cognitive frames may be used to describe meanings of predicates and sentences in both of the resources.

The organization of the rest of the paper is as follows. In Section 2, backgrounds to FN, JFN, and KCF will be discussed. Section 3 deals with the methodology we adopted. Section 4 discusses the experimental settings and the preliminary results. It will be shown that according to the accuracy of crowdworkers responses, the predicates used in our experiments can be classified into three categories. Section 5 gives conclusions and prospects.

2. Related Work

FN is based on the framework of Frame Semantics. Cognitive frames correspond to word meanings.[2] Each cognitive frame has its own frame elements (FEs), similar to semantic roles in other theories, except that FEs are specific to each cognitive frame. The `Sending` frame ("a SENDER plans the PATH of a THEME and places it in circumstances such that it travels along this PATH under the power of some entity other than the SENDER") is an example of a cognitive frame and SENDER, PATH, and THEME are its FEs. LUs are a pairing of a lemma with a meaning, i.e., with a cognitive frame. For example, the English lemma *express* has at least two distinct LUs, namely, `Sending`.*express*.v and `Encoding`.*express*.v. That is, the verb *express* may be used to mean "to send in the post with a short delivery time," that is, with the meaning of the `Sending` frame. In addition, the same lemma *express* may also be used in a situation in which "a PERSON encodes a MESSAGE or mental content, broadly understood, in a particular MANNER," that is, in the `Encoding` frame.

The FrameNet database contains: definitions of cognitive frames and of their FEs; annotated corpus example sentences of LUs; and **valence patterns** (combinatorial possibilities of arguments and adjuncts, in terms of FEs, phrase types (PTs), and grammatical functions (GFs)) of LUs (cf. Table 1). For example, the English LU `Sending`.*express*.v currently has 39 valence patterns in FN, including "[SENDER.NP.Ext] *send* [THEME.NP.Obj] [PURPOSE.VP*to*.Dep]"[3] as in "[<SENDER> member states of the Arab League] *sent* [<THEME> troops] [<PURPOSE> to help the Palestinian Arabs]."

JFN is compatible with FN: sharing definitions of cognitive frames and their FEs, database structures, methodologies and some of the tools (Ohara, 2014). As shown in Table 1, there are currently 5029 LUs in JFN, consisting of: 1136 verbs, 132 adjectives, 152 adjectival nouns, and 3307 nouns.

There have been studies that assign FN cognitive frames to sentences using crowdsourcing (Hong and Baker, 2011; Fossati et al., 2013; Chang et al., 2015). Their methods basically present crowdworkers with example sentences or simplified frame definitions and ask them to select one from several choices. Unlike previous studies, our method involves not only word sense disambiguation (cognitive-frame disambiguation) but also linking two different types of frame knowledge, namely, JFN (i.e., top-down frame knowledge) and KCF (i.e., bottom-up frame knowledge). Moreover, as will be discussed below, our crowdsourced task involves example sentence selection and thus requires no prior knowledge of Frame Semantics on the part of crowdworkers.

In KCF, each case frame is represented as a predicate and a set of its case slots (or case markers) with their instance words. KCF contains verbs, copulas, adjectives and adjectival nouns, but not nouns. Table 2 is a partial list of the case frames of the verb *okuru* 'send' in KCF.

KCF Case Frame ID	Case Slots	Instance Words
okuru (1)	*ga* (NOM[4]) *o* (ACC) *ni* (DAT)	*watashi* 'I':374, ... *meeru* 'mail':211755, ... *keitai* 'cell phone':30944, ...
okuru (2)	*ga* (NOM) *o* (ACC) *ni* (DAT)	*josei* 'women':489,... eeru 'yell':70314, ... *senshu* 'athlete':3478, ...
okuru (3)	*ga* (NOM) *o* (ACC) *ni* (DAT)	*watashi* 'I': 125, ... *shinsei* 'application': 35477, ... *kaisha* 'company': 1367, ...
...	...	...

Table 2: Examples of KCF case frames for the predicate *okuru* 'send.' The numbers denote frequencies.

Here, the case frame *okuru* (1) consists of: the case slot *ga* followed by its instance words *watashi* 'I,' *dare* 'who,' *hito* 'person' ...; the case slot *o* followed by *meeru* 'mail,' *messeeji* 'message' ...; and the case slot *ni* followed by

[1] We would like to thank an anonymous reviewer for pointing this out to us.

[2] In Frame Semantics literature, terms such as cognitive frames (Fillmore, 1982, p.117 (Geeraerts (Ed.), 2006, p.379)), Fillmore and Baker, 2010, p. 314), semantic frames (Ruppenhofer et al., 2010), linguistic frames (Fillmore and Baker, 2010, p.338) and frames (Fillmore and Baker, 2010, p.314) have been used to refer to the same notion. In this paper, in order to distinguish the notion from case frames, we will use "cognitive frames."

[3] NP: noun phrase, VP: verb phrase, Ext: External Argument (i.e., Subject), Obj: Direct Object, Dep: Dependent (i.e., anything other than subject and direct object)

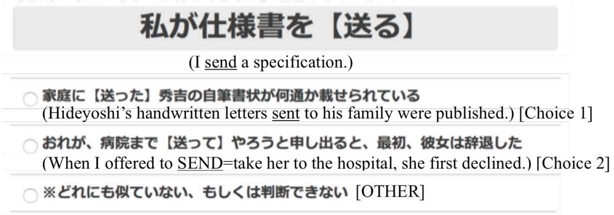
[4] "NOM" stands for the nominative case; "ACC" the accusative; and "DAT" the dative.

keitai ‘cell phone,’ *hito* ‘person,’ *tomodachi* ‘friend.’ Here, even though the three case frames of *okuru,* namely, *okuru* (1) through *okuru* (3), contain the same set of case slots *ga* (the nominative), *o* (the accusative), and *ni* (the dative), the instance words that each of the case slots accompanies are different. In other words, each case frame in KCF represents a “usage” of a predicate. The number of KCF case frames of a predicate usually exceeds the number of JFN LUs of the same predicate (in other words, exceeds the number of JFN cognitive frames that the predicate is associated with), it may be possible to say that a “usage” that each KCF case frame represents is more fine-grained than a “meaning” that a JFN cognitive frame represents. Unlike JFN valence patterns, however, KCF case frames do not at all include semantic information about case filler words. That KCF does not contain semantic information at all means that it does not have semantic information incompatible with JFN. Also, JFN cognitive frames can be used to describe meanings of predicates in KCF. It is thus possible to link JFN annotated sentences with KCF case frame example sentences via cognitive frames.

We use the latest version of KCF, which was constructed by applying Chinese Restaurant Process-based clustering (Kawahara and Kurohashi, 2006; Kawahara et al., 2014) to 10 billion Japanese sentences. KCF has about 110,000 predicates and 5.4 case frames on average for each predicate.

3. Methods

We link each KCF case frame of a predicate with one of the JFN cognitive frames that corresponds to the same meaning of the predicate. We cast this linking process as a crowdsourced task of example sentence selection. Figure 1 shows a screenshot of the crowdsourced sentence selection task for the case frame (3) of *okuru* ‘send’ in Table 2.


私が仕様書を【送る】

(I send a specification.)

○ 家庭に【送った】秀吉の自筆書状が何通か載せられている
(Hideyoshi’s handwritten letters sent to his family were published.) [Choice 1]

○ おれが、病院まで【送って】やろうと申し出ると、最初、彼女は辞退した
(When I offered to SEND=take her to the hospital, she first declined.) [Choice 2]

○ ※どれにも似ていない、もしくは判断できない [OTHER]

Figure 1: An example of crowdsourced sentence selection tasks

An example sentence for the case frame (3) from KCF was presented to crowdworkers and they were asked to select a JFN example sentence that is most similar to the presented sentence. The first two choices in Figure 1 are example sentences in JFN for `Sending`.*okuru*.v (Choice 1) and for `Bringing`.*okuru*.v (Choice 2) respectively. In addition to these two choices, we made another choice “No similar sentences exist or impossible to judge” (“OTHER”, hereafter), which is to be selected if the presented example sentence from KCF is not similar to either of the JFN example sentences or if it is impossible to judge from the presented sentence. We hypothesized that when “OTHER” was selected by many, there might be something to re-examine in the cognitive-frame assignment for the predicate in JFN (cf. Section 4.2).

We assumed that sentences shown to crowdworkers (both the presented sentence and the Choice 1 and Choice 2 sentences) should be short, so that it would be easy for them to understand their meanings. To generate such a sentence for each case frame in KCF, we selected a sentence that had the highest generative probability based on a language model from the set of example sentences that constitute the target case frame. By this method, we were able to select a sentence that was short and easy to understand. We adopted an RNN language model (Mikolov et al., 2010) to calculate the generative probability of a sentence. This RNN language model was trained on a web corpus consisting of 10 million Japanese sentences.

To generate an example sentence for a JFN cognitive frame, we manually selected the shortest example sentence from the set of example sentences that belong to the target JFN cognitive frame. The reason why we picked the shortest example sentences was to take into account the screen sizes of PCs and of smart phones and to make it easier for crowdworkers to read them. If a selected example sentence was longer than 60 characters, it was shortened by hand.

4. Experiments

4.1 Experimental Settings

There are currently 935 predicates that exist both in JFN and KCF. Among these predicates, we conducted experiments on 37 predicates (27 verbs, 5 adjectives and 5 adjectival nouns) that have two JFN cognitive frames and at least one example sentence for each in the JFN database. There were only 37 predicates that met the criteria above. These 37 predicates have 712 case frames in total in KCF.

The predicates that exist only in JFN are mostly complex prepositions (e.g. *ni_kansuru* ‘with respect to’) and compound nouns that may also be used as verb stems (e.g. *syookyaku_shori* ‘incineration’), which are not included in KCF.

There are approximately 110 thousand predicates that exist in KCF but not in JFN. This is because KCF distinguishes predicates with auxiliaries that cause case alternations. For example, in addition to *uru* ‘sell,’ a “bare” predicate, KCF has additional separate predicates with the same stem and an auxiliary verb beginning with *–te*, such as *ut-teiru, ut-tekuru, ut-tekureru*. There are 50 thousand predicates with a *–te* auxiliary verb in KCF. Furthermore, KCF distinguishes predicates with passivizing and causativizing suffixes, from the active predicates without such suffixes. There are 31500 predicates without passivizing/causativizing suffixes; 5300 predicates with the passivizing suffix *-(r)are*; and 1700 predicates with the causativizing suffix *-(s)ase.* In contrast, in JFN, uses of predicates with a *–te* auxiliary verb and uses of predicates with the passivizing/causativizing suffix are included in the same LUs and case alternations are recorded as different valence patterns of the same LUs (cf.

Section 2). KCF also contains many infrequent predicates (e.g. *nyuuzan_suru* 'go into a mountain,' *nikusyoku_da* 'carnivorous') that JFN does not contain.

We employed Yahoo! Crowdsourcing[5] to crowdsource the linking task. We asked 10 crowdworkers for the linking task of each case frame. Their answers were aggregated by majority voting. To alleviate the influence of malicious crowdworkers, we used gold questions, i.e., easy questions to which we had known the correct answers beforehand. We eliminated the crowdworkers who had not correctly answered the gold questions. As a result, in total 272 crowdworkers participated in the task, and it took approximately two hours to complete the task. The total cost was approximately 25,000 JPY.

4.2 Results and Discussions

We examined the responses of the crowdworkers for each case frame of each predicate, by manually checking whether their responses were correct or not. Specifically, we analyzed whether the JFN cognitive frame that got the largest number of votes was correct or not. Two JFN annotators evaluated the results of the crowdsourced task. After each of the two annotators individually evaluated the results, the principal JFN annotator compared the two evaluations (one by herself and another by the other JFN annotator) and gave the final evaluation.[6] There were inter-annotator agreements for the majority of the sentences.

KCF Case Frame ID	KCF Target Sentence	JFN Cognitive Frame with the largest # of votes
okuru (1)	*watashi tachi ga iimeeru o okutta* (We sent email)	✔Sending
okuru (2)	*futari ga seien o okuru* (Two people SEND cheers)	✔"OTHER"
okuru (3)	*watashi ga shiyoosho o okuru* (I send a specification)	✔Sending
okuru (4)	*futari ga setsuyaku seikatsu o okuru* (Two people SEND=live frugal lives)	✔"OTHER"
okuru (5)	*watashi ga senga o okuru* (I send specification)	✔Sending
okuru (6)	*watashi ga tookyoo eki made sannin o okuru* (I SEND =take three people to Tokyo Station)	✔Bringing
okuru (7)	*jibun ga seishun jidai o okutta* (I SENT=spent [my] youth)	✔"OTHER"
okuru (8)	*watashi ga fakkusu de okuri mashoo ka* (Shall I send [it] by fax?)	✔Sending
okuru (9)	*boku no noo ga kiken shingoo o okuru* (My brain sends a danger signal)	×OTHER

Table 3: Results of JFN cognitive frame assignments to KCF case frames for *okuru* 'send' by crowdworkers[7]

Table 3 shows the result of evaluating the responses by crowdworkers for all the 9 case frames of the verb *okuru* 'send.' There were varying degrees of accuracy depending on the predicate. After evaluating the responses by the crowdworkers, we classified the 37 predicates into three categories based on two factors: the sematic closeness of the two relevant JFN cognitive frames; and whether the two JFN cognitive frames actually characterize the meanings of the predicate in question. The proposed three categories of predicates are the following:

Category I: None of the criteria for Categories II or III below applies. That is, the two JFN cognitive frames, which represent the two meanings of the predicate, are semantically distinct.

> e.g. *okuru* 'send' (The Sending frame, in which a SENDER does not travel with a THEME, is semantically distinct from the Bringing frame, in which an AGENT travels together with a THEME.)

Category II: The two meanings of the predicate are semantically close. There are two cases: the two JFN cognitive frames are related via JFN frame-to-frame relations; or not.

An example of the former is *iku* 'go.'

> e.g. *iku* 'go' (The Motion and Self_motion frames differ only in whether the entity that moves is a living being or not and the two cognitive frames are linked to each other via the Inheritance frame-to-frame relation.)

An example of the latter is *kaku* 'write.'

> e.g. *kaku* 'write' (The Text_creation frame, having to do with creating a TEXT that contains meaningful linguistic tokens, and the Spelling_and_pronouncing frame, pertaining to realizing a SIGN in some FORMAL_REALIZATION, are semantically close to each other but they are not related by any frame-to-frame relation.)

Category III: The cognitive frames assigned to the predicate in JFN do not correctly characterize its meanings. There are two cases: the predicate was incorrectly assigned a cognitive frame in JFN; or the predicate by itself (that is, not as a support predicate that accompanies a specific noun phrase) evokes another cognitive frame that has not been assigned to the predicate in JFN.

An example of the former is *tekisetsu-da* 'appropriate.'

> e.g. *tekisetsu-da* 'appropriate' (The Suitability and Desirability frames were assigned to this predicate in JFN. However, as the latter cognitive frame has to do with an EVALUEE being judged for its quality, i.e. how much it would be probably liked, it does not characterize the meaning of the predicate and thus should not have been assigned to the predicate in JFN.)

An example of the latter is *ataeru* 'give.'

> e.g. *ataeru* 'give' (In addition to the Giving and Supply frames that have been assigned to the verb in JFN, the Objective_influence and Subjective_influence frames should also be assigned to it.)

[5] https://crowdsourcing.yahoo.co.jp/

[6] The two annotators are both experienced JFN annotators, but whereas the principal annotator is formally trained in linguistics and in the theory of Frame Semantics, the other annotator is not.

[7] The symbols ✔ (correct) and × (incorrect) show the evaluation by JFN annotators.

Table 4 summarizes the accuracy of the crowdworkers' responses for each of the three categories of the predicates. It shows that the predicates in Category I, namely, those having two semantically distinct meanings, achieved the highest accuracy. Category II predicates, with two semantically close meanings, followed Category I predicates in the accuracy. Category III predicates, with incorrect or incomplete cognitive-frame assignments, had the lowest accuracy.

Predicate Category	# of Predicates	Accuracy
I	9	83.9%
II	11	57.9%
III	17	26.3%

Table 4: Micro Average of Accuracy

Our tentative hypotheses include the following:

1) When the two JFN cognitive frames assigned to a predicate are semantically close, it is difficult for crowdworkers to correctly distinguish between the two meanings (Category II);
2) When the assignment of a cognitive frame in the JFN database is incorrect, it is difficult for crowdworkers to make a distinction among the "correct" word meanings of the predicate (Categories III);[8]
3) When the predicate involves more than two meanings, it is difficult for crowdworkers to correctly make a distinction among them (Category III)

There are other possible causes for crowdworkers' mistakes. Some of the KCF sentences we presented to crowdworkers did not include all the syntactic arguments (i.e., all the case slots) and consequently the sentences were vague. It was thus impossible for crowdworkers to determine their meanings. In our future experiments we plan to use sentences with all the case slots filled.

Also, there are sentences in which a predicate constitutes a part of an idiom.[9] With such sentences, judgments by crowdworkers varied. Examples include:

(1) *chie* *o* shiboru 'rack one's brains'
wisdom ACC squeeze

(The whole phrase evokes the `Cogitation` frame.)

(2) *me* *o* toosu 'skim through'
eye ACC pass

(The whole phrase evokes the `Reading_perception` frame.)

(3) *sode* *o* toosu 'put on a shirt'
sleeve ACC pass

(The whole phrase evokes the `Drssing` frame.)

(4) *hooan* *o* toosu 'pass (a bill)'
bill ACC pass

(The whole phrase evokes the `Successfully_communicate_message` frame.)

Turning to the "OTHER" option in the crowdsourced task, it appears that crowdworkers resorted to this option when a support-predicate usage was involved in the presented sentence. For example, in Table 3, case frame (2) (*seien o okuru*: literally 'send cheers,' in other words, 'cheer'), case frame (4) (*seikatsu o okuru*: literally 'send a life,' in other words, 'live'), and case frame (7) (*seishun jidai o okuru*: literally 'send youth,' in other words, 'spend (one's) youth') involve such uses of the verb *okuru*. Therefore, the "OTHER" option may be used as a clue to finding support-predicate uses of predicates. We hope to investigate crowdworkers' uses of the "OTHER" option further.

5. Conclusion and Prospects

We proposed a method to crowdsource the assignment of JFN cognitive frames, that is, word meanings, to KCF case frames, by matching an example sentence from KCF and another from JFN that share the same meaning of a predicate. Our initial experiments with predicates that have two JFN cognitive frames yielded promising results, especially with regard to Category I predicates, namely, those having two meanings that are not semantically close to each other.

Our next step is to conduct experiments with the remaining 898 predicates that have been assigned three or more cognitive frames in JFN. For this task, we plan to use the frame-to-frame relations in JFN.

Although our ultimate goals include scaling up the size of annotated sentences in JFN, so far we have concentrated on the task of cognitive-frame disambiguation. The whole FN/JFN annotation process also involves assignments of FEs and thus our longer-term goals include assigning JFN FEs to individual case slots (i.e., case-marked NPs) of each KCF case frame as well. We estimate this task to be relatively easy for crowdworkers compared to the task reported in this paper, that is, compared to finding a JFN annotated sentence similar to an example sentence of a KCF case frame.

Furthermore, in order to increase the coverage of JFN, we plan to work on predicates that exist in KCF but not in JFN, by mapping each KCF case frame to a JFN cognitive frame. We will first focus on the "bare" predicates, which do not have passivizing/causativizing suffix or a *–te* auxiliary, in KCF.

Bibliographical References

Chang, N., Paritosh, P., Huynh, D., and Baker, C. (2010). Scaling semantic frame annotation. In Proceedings of the 9th Linguistic Annotation Workshop, pages 1–10,

[8] We have yet to investigate whether correcting the assignment of JFN cognitive frames for these predicates would indeed improve the accuracy of responses by crowdworkers.

[9] In addition to having the "OTHER" option, it might be possible to add another choice of "IDIOM." However, following guidelines for crowdsourcing, we decided to keep each individual task as simple as possible for crowdworkers and thus did not make a choice of "IDIOM."

Denver, Colorado, USA, June. Association for Computational Linguistics (ACL).

Fillmore, C.J., and Baker, C.F. (2010). A frames approach to semantic analysis. In B. Heine and H. Narrog (Eds.), *The Oxford Handbook of Linguistic Analysis*. Oxford: Oxford University Press, pp. 313–339.

Fossati, M., Giuliano, C., and Tonelli, S. (2013). Outsourcing framenet to the crowd. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 742–747, Sofia, Bulgaria, August. Association for Computational Linguistics (ACL).

Hong, J. and Baker, C.F. (2011). How good is the crowd at ”real” wsd? In Proceedings of the 5th Lin- guistic Annotation Workshop, pages 30–37, Portland, Oregon, USA, June. Association for Computational Linguistics (ACL).

Kawahara, D. and Kurohashi, S. (2006). Case frame compilation from the web using high-performance computing. In Proceedings of the Fifth International Conference on Language Resources and Evaluation, pages 1344–1347, Genoa, Italy, May. European Language Resource Association (ELRA).

Kawahara, D., Peterson, D., Popescu, O., and Palmer, M. (2014). Inducing example-based semantic frames from a massive amount of verb uses. In Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, pages 58–67, Gothenburg, Sweden, April.

Mikolov, T., Karafiát, M., Burget, L., Cernockỳ, J., and Khudanpur, S. (2010). Recurrent neural network based language model. In Proceedings of Interspeech 2010, pages 1045–1048, Makuhari, Japan, September.

Ohara, K. (2014). Relating frames and constructions in Japanese FrameNet. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC’14), pages 2474–2477, Reykjavik, Iceland, May. European Language Resource Assocation (ELRA).

Ruppenhofer, J., Ellsworth, M., Petruck, M.R.L., Johnson, C.R., Baker, C.F., and Scheffczyk, J. (2016). FrameNet II: Extended theory and practice, Revised November 1. https://framenet.icsi.berkeley.edu/fndrupal/the_book

Language Resource References

Kyoto University Case Frames. http://www.gsk.or.jp/en/catalog/gsk2008-b/

Japanese FrameNet. http://jfn.st.hc.keio.ac.jp/

The Multilingual FrameNet Shared Annotation Task: a Preliminary Report

Tiago Timponi Torrent[1], Michael Ellsworth[2], Collin Baker[2], Ely Edison da Silva Matos[1]
[1]Federal University of Juiz de Fora - FrameNet Brasil, Juiz de Fora, MG - Brazil
[2]International Computer Science Institute, Berkeley, CA - USA
{tiago.torrent, ely.matos}@ufjf.edu.br, {infinity, collinb}@icsi.berkeley.edu

Abstract

This paper presents the shared annotation task devised by the Multilingual FrameNet project together with partner projects. The shared framenet annotation task intends to probe how comparable frames are across languages by annotating translated and comparable texts using the same semantic standards in multiple languages. This paper reports on the initial work of agreeing on annotation standards, building annotation tools, and the results from the first joint frame annotations, from a TED talk and its translation into Brazilian Portuguese. The results indicate that the joint annotation task is feasible with existing FrameNet frames: over 80% of frame-bearing words in the Brazilian Portuguese translation of the TED talk fit precisely in frames found in Berkeley FrameNet's release 1.7. However, even languages as typologically similar as English and Brazilian Portuguese show some differences in density of frame-bearing words and the frequency of frame-bearing words by part-of-speech.

Keywords: Multilingual FrameNet, Shared Annotation, Interlingual Comparison

1. Multilingual FrameNet

Since 1997, the FrameNet Project at the International Computer Science Institute, in Berkeley, California, has been building a richly detailed lexical database of the core vocabulary of contemporary English, implementing the theory of Frame Semantics, developed by the late Prof. Charles Fillmore and colleagues (Fillmore 1976, 1982, Fillmore & Baker 2010). The Berkeley project has defined semantic frames, frame elements (roles) in these frames, and lexical units (word senses) which evoke the frames, extracted text from corpora and annotated the instances of these lexical units in the texts. The Berkeley FrameNet lexical database (browsable at http://framenet.icsi.berkeley.edu) currently contains 1,224 semantic frames, each of which has an average of 9.7 frame elements (FEs), and comprises 13,639 lexical units (LUs). There are 202,229 manually annotated instances of these lexical units, each containing annotation of the FEs that appear in the sentence.

All of this research has been done on English, but the researchers have frequently considered the obvious question: to what extent are the semantic frames created for English appropriate for analyzing other languages. Fortunately, inspired by the work at ICSI, a number of related projects have been developing frame semantic lexical databases for roughly a dozen languages, which vary in size, methodology, and availability. In all cases, the new projects have taken the Berkeley (English) frames as a starting point, although some have adhered more closely to the example of English. In general, these projects have found that a large proportion of the target-language words fit comfortably in those frames. The FrameNet team has now embarked on a Multilingual FrameNet project, developing alignments across many of these FrameNets, seeking a better understanding of cross-linguistic similarities and differences in frame structure. Alignment on the frame level is often quite easy, as many projects have kept names or ID numbers which refer to the Berkeley frames. Going beyond frame connections, other techniques are being used to cluster and align lexical units across languages. One of these is using multilingual word vectors (Hermann & Blunsom 2014) which can be computed for a large range of languages from a wide variety of texts, and (unlike, e.g. bilingual dictionaries) lend themselves to quantitative measures of goodness of fit. We are currently testing these, but also considering techniques based on other curated resources, such as Open Multilingual WordNet (Bond & Foster 2013) and BabelNet (Navigli and Ponzetto 2012).

2. The Shared Annotation Task

The shared annotation task was devised in part as a means to evaluate the complexity of the work required to align the FrameNets developed for different languages during the past decade and more. By annotating either translations of a given text or comparable texts from the same genre and on the same topic, we aim to assess what kinds of differences must exist between FrameNets for different languages in order to provide an adequate analysis of the lexicon of each language. Moreover, the shared annotation task will generate a collection of texts annotated with frames and LUs for several languages, which can be used in the future, for instance, as training data in a variety of applications.

In the shared annotation task, annotators were limited to using the frames and frame elements from the 1.7 release of the Berkeley FrameNet data (BFN 1.7), so that everyone would annotate on the same basis. We anticipated that in many cases, a BFN 1.7 frame would be the best-fitting frame (BFF) for a word in another language, but in other cases, it might not be, suggesting that different languages might require different adaptations to those frames. In the latter case, annotators are instructed add the LU to the nearest BFN 1.7 frame, but also to indicate why that is not the best-fit frame for the LU. They could choose among the following predefined categories, or “other”:

- **Different Perspective:** the LU imposes a perspective that is different from the one in the original frame.
- **Different Causative Alternation:** the LU requires a causative interpretation that is not present in the original frame, which may be either inchoative or stative.
- **Different Inchoative Alternation:** the LU requires an inchoative interpretation that is not present in the original frame, which may be either causative or stative.
- **Different Stative Alternation:** the LU requires a stative interpretation that is not present in the original frame, which may be either causative or inchoative.
- **Too Specific:** the LU requires a frame more generic than the one available in the original database.
- **Too Generic:** the LU requires a frame more specific than the one available in the original database.
- **Different Entailment:** the LU has different entailments than the ones afforded by the original frame.
- **Different Coreness Status:** some non-core FE should be core in the target language.
- **Missing FE:** there should be a FE in the original frame that is missing.
- **Other:** all other non-listed cases.

Each annotation must include, at least, the Frame Element, Grammatical Function and Phrase Type layers. Labels in each layer can be tailored to the specific needs of each language, and, new layers can be added to the annotation.

These policies on the shared task were then carried out in a first round of shared annotation on a translated text, described in Section 2.1, using a web annotation tool developed by FrameNet Brasil, described in Section 2.2.

2.1. The Text

The first text to be annotated in the shared annotation task is the transcription of the TED Talk "Do Schools Kill Creativity?" (Robinson 2006). This is currently the most frequently viewed TED Talk, with more than 49 million views. The transcription of the 20-minute talk in English contains 267 sentences. This transcription has been translated to 61 languages by TED community members; the Brazilian Portuguese version, which will be discussed below, has 271 sentences.

2.2. The Annotation Tool

The shared annotation task is carried out with the FrameNet Brasil WebTool 3.0: a web-based database management and annotation tool, designed to allow easy customization of layers and labels from a multilingual perspective (Matos & Torrent 2016).

Because it is web-based, the tool does not require the annotation teams to install any software. Moreover, it allows teams to create language-specific annotation labels for Grammatical Functions, Phrase Types and other information. Annotators can even add new layers to the annotation system if necessary, directly in the tool interface, without having administrator privileges. This flexibility enables teams to create the analytical categories they need to address the specifics of their languages.

3. Preliminary Report

So far, consistent annotations of the TED Talk have been made for English (2 annotators) and Brazilian Portuguese (7 annotators). In this paper, we offer a preliminary contrastive report on those annotations, based, on the first 30 sentences of text in both languages, which we will refer to those sentences as the **sample text**.

The sample text comprises a total of 425 words for English and 322 for Brazilian Portuguese. Among those words, 89 different LUs were identified for English, yelding 132 annotation sets. (Each instance of each LU constitutes a separate annotation set.) For Brazilian Portuguese, 107 different LUs were identified, yelding 146 annotation sets. The annotation set/word ratio is then 0.31 for English, and 0.45 for Brazilian Portuguese. The density of annotation in the English sample text compares to 0.17 for all the full text annotation in Berkeley FrameNet; this may be due in part to a more complete annotation of the sample text and in part to a greater density of frames in the spoken genre. The difference in the density across languages is shown in more detail in Table 1, which gives the distribution of annotation sets by POS in each language.

Some of the differences, especially for conjunctions, stem from differences between the projects as to which parts of the semantics should be represented by FrameNet lexical annotation and which parts should be represented by constructions. Note that there is very little difference between the languages w.r.t. the density of annotation of verbs; we suspect that there may be two reasons for this:

1. Verbs tend to be the main predicates in sentences, evoking the central eventive frames, so translations might tend to keep the same number of central eventive frames.

2. Because semantic frames are arguably better models for events than for entities, FrameNet may simply have better, more robust models for events, which tend to be expressed more often by verbs in both languages.

POS	English	Br-Portuguese
Adjective	16	26
Adverb	6	11
Conjunction	8	20
Noun	48	51
Number	4	3
Preposition	9	5
Pronoun	-	2
Verb	41	40
TOTAL	**132**	**148**

Table 1: Distribution of annotation sets in the TED Talk sample text by part of speech of LU in each language.

In order to gauge the similarity between the annotations for English and for Brazilian Portuguese, a similarity score was calculated for each aligned pair of sentences, based on the frames evoked by the LUs in each language.

First we found the total number of frames evoked in each sentence. (When the total was different between the two languages, we used whichever number was higher.) Then the number of frames that were the same in both languages was counted and that number was divided by the total. For example, there were a total of 9 frames in sentence 7, and 4 of them were the same across languages, so the similarity score is 4/7, or 0.44. Table 2 presents the similarity scores for each of the 30 sentence pairs in the sample text and the average for all of them.

In Table 2, sentence pairs 1, 2 and 13 are marked with "N/A” because no frames have been assigned to these sentences in either language. They consist of two greetings (pairs 1 and 2) and one tag question (13). There are a number of cases where the similarity score is low because both annotation teams added an LU that was not a perfect fit to a frame from BFN 1.7, but they each chose a different best-fit frame. We have treated these like “normal” cross-linguistic differences, but some other treatment might be appropriate.

Pair	Total Frames	Equal Frames	Score
1	N/A	N/A	N/A
2	N/A	N/A	N/A
3	1	1	1.00
4	4	1	0.25
5	1	1	1.00
6	7	4	0.57
7	9	4	0.44
8	2	1	0.50
9	10	5	0.50
10	3	1	0.33
11	2	1	0.50
12	4	2	0.50
13	N/A	N/A	N/A
14	3	3	1.00
15	6	5	0.83
16	7	2	0.29
17	4	1	0.25
18	5	0	0.00
19	2	1	0.50
20	18	5	0.28
21	5	1	0.20
22	5	3	0.60
23	11	5	0.46
24	7	4	0.57
25	11	5	0.46
26	8	8	1.00
27	13	6	0.46
28	5	2	0.40
29	3	1	0.33
30	11	5	0.45
Average Frame Similarity Score			**0.51**

Table 2: Frame Similarity Score between Languages per sentence pair.

In the following two sections, we discuss the main issues that emerged during the annotation for each language. Section 3.3 provides some cross-linguistic comparison of annotated sentences.

3.1. The Annotation for English

Annotating the TED talk has been challenging for Berkeley FrameNet, since it is a spoken genre, with a large number of conversation-specific LUs and constructions, such as *you know, ...* and *I mean….* However, for the rest of the lexical items in the text, it has been possible to use the frames of BFN 1.7 without modification in the vast majority of instances. Out of 132 total LU instances, 125 (95%) fit their frame perfectly, 5 (e.g., *creativity.n*, *blood.n*) were in only found in frames that were too generic for the use in this text, 1 (*curiously.adv*) was in a frame belonged to a different perspective, and 1 (*interest.n*) should actually be a MWE (*vested interest.n*), evoking a frame that does not exist in BFN 1.7.

However, these numbers hide a policy difference in the annotation of the English text compared with the Brazilian Portuguese. Until now, Berkeley FrameNet has considered pure conjunctions (e.g., *and.c*) and conversationally-grounded items like *actually.adv* and *you know.v* to be outside the scope of BFN annotation, since they are so entangled with interactional frames that FrameNet has not yet defined and with non-lexical constructions. There are 10 instances of *and.c* in the sample text, and 11 conversational particles, all of which would belong to very poorly fitting frames. If these are considered, then only 82% of LU instances belong to an appropriate frame in the annotation of the English text, which is remarkably similar to the ratio for Brazilian Portuguese, as we will see in the next section.

3.2. The Annotation for Brazilian Portuguese

Besides issues related to the fact that the TED Talk is a spoken genre, as pointed out in 3.1, the annotation of the sample text for Brazilian Portuguese was expected to pose additional challenges due to the way the shared task was designed. Since no changes could be made to BFN 1.7, we anticipated that there would be many cases in which an LU appearing in the text would have to be created in a non-BFF frame, and we provided means for annotators to do this, and save an explanation of why the frame chosen is not ideal, as a suggestion for someone about how to define the proper frame later.

Reason	Count
Different Perspective	1
Too Generic	5
Different Entailment	1
Different Coreness Status	1
Missing FE	4
Other	8
TOTAL	**20**

Table 3: Reasons for creating LUs with non-BFF status in Brazilian Portuguese

However, this turned out to be not very common. Among the 107 different LUs in the Brazilian Portuguese text, only 20 (18.7%) were created in non-BFF frames, meaning that Berkeley FrameNet frames provided an adequate model for more than 80% of the Brazilian Portuguese LUs. Moreover, if one considers the reasons behind the non-BFF status (shown in Table 3), BFN 1.7 frames seem to be even more easily expandable into Brazilian Portuguese. The "Too Generic" cases, representing one fourth of the LUs created in non-BFFs, indicate that the usage would require a new, more specific frame not yet available in BFN 1.7; this proposed new frame would inherit from the non-BFF frame in which the LU in the text was created. Examples are LUs like *deus.n 'god'*, in the Entity frame, and *e.c 'and'* in the Relation frame. The "Missing FE" cases all refer to non-core FEs which could be easily added to the frames, even in English, such as a Condition FE in the Concessive frame, and a Degree FE in the Causation frame. Some of the "Other" cases, however, refer to more complex (and interesting) cases, which will be discussed in the next section.

3.3. Some cross-linguistic examples

As it can be seen from Table 2, cross-linguistic frame similarity scores vary considerably from sentence pair to sentence pair. In this section, we provide examples covering three different parts in this range: sentence pairs with a 1.000 similarity score, sentence pairs with low similarity scores due to the occurence of non-BFF frames, and sentence pairs with similarity scores close to the average, which are due to differences in translation and/or language structure.

The high end of the range is exemplified by sentence pair 26, in which sentences (1) and (2) were annotated for the same 8 frames in each language:

(1) If you think of it, children starting school this year will be retiring in 2065.

(2) Se formos pensar, as
if go.FUT.SUBJ.1PL think.INF the
crianças entrando na escola
children enter.PTCP in the school
esse ano estarão se
this year be.FUT.3PL them-RFL
aposentando em 2065.
retire.PTCP in 2065

Table 4 presents the 8 frames selected for annotation and the LUs evoking each of them in English and Brazilian Portuguese.

As it can be seen from Table 4, LUs evoking the frames in both languages have the same POS. Also, none of them was assigned the non-BFF type. Although there are structural differences in the translation of (1) into (2) - e.g. the fact that think.v takes a second person subject in English, while pensar.v takes a first person plural subject in Brazilian Portuguese - such differences do not concern frame evoking material. Three other sentence pairs received a score of 1,000, two have one LU for each language, and the other has 3.

Frame	En LU	Br-Pt LU
Conditional_occurence	if.c	se.c
Cogitation	think.v	pensar.v
People_by_age	child.n	criança.n
Activity_start	start.v	entrar.v
Locale_by_use	school.n	escola.n
Calendric_unit	year.n	ano.n
Quitting	retire.v	aposentar.v
Temporal_collocation	in.prep	em.prep

Table 4: Frames for which sentences (1-2) were annotated and LUs evoking them in each language.

On the low end of the similarity score scale, with a score of 0.00, we find sentences (3) and (4) in pair 18.

(3) And you're never asked back, curiously.

(4) E curiosamente ninguém te
and curiously no one you
convida de novo.
invite.PRES.3SG again

Table 5 shows the LUs annotated in each language and the frames they evoke. Note that there are no corresponding frames between the two languages. A "---" indicates that the frame was not evoked in one of the languages.

Frame	En LU	Br-Pt LU
Frequency	never.adv	---
Request	ask.v	---
Locative_relation	back.adv	---
Typicality	curiously.adv	---
Relation	---	e.c
Manner	---	curiosamente.adv
People	---	ninguém.n
Have_visitor_over	---	convidar.v
Event_instance	---	de novo.adv

Table 5: Frames for which sentences (3-4) were annotated and LUs evoking them in each language.

The low score in this sentence pair illustrate how different choices for non-BFF frames impact the comparability between the original sentence and its translation in terms of semantic frames. The English sentence has one LU created in a non-BFF frame (*curiously.adv*), which should actually be handled as a sentence-level modifier; it ironically suggests that the hearer should understand why educators are seldom asked again by the same host. Such a frame, which invokes the full conversational context, has not yet been defined for either language. In the Brazilian Portuguese translation, three LUs were created in non-BFF frames. One of them, *curiosamente.adv* - which actually translates as *curiously.adv* - was created in the Manner frame, which is too generic and includes LUs such as *manner.n* and *way.n*, but not adverbs actually indicating manner. This use of Portuguese *curiosamente.adv* should probably be handled like English *curiously.adv.*

The LU *convidar.v* was created in a non-BFF frame for two reasons: first, because there was a missing non-core FE, Particular_iteration, and, second, because the Have_visitor_over frame seems to be, in fact, preceded by the frame evoked by *convidar.v*.

The LU *e.c*, which translates as *and.c* was created in the Relation frame, a very generic frame not really used by BFN for conjunctions such as this, as pointed out in 3.1.

In the middle of the score continuum, sentence pair 25, with a score of 0.46, has 5 coincidental frames out of 11. The sentences of this pair are shown in (5) and (6).

(5) We have a huge vested interest in it, partly because it's education that's meant to take us into this future that we can't grasp.

(6) Nos interessamos tanto por
us-RFL be-interested.PRES.1PL so much for
ela em parte porque é da
she in part because be.PRES.3SG of
educação o papel de nos
education the role of us
conduzir a esse futuro misterioso.
conduct.INF to this future misterious

Table 6 shows the frames evoked by the LUs in this sentence pair.

Differences between the frames evoked by the LUs in each version of this sentence can be classified into two types: (a) structural differences in the predicates and (b) the cascade effect of those on their modifiers.

The first predicate in each sentence is that encoding the interest people have in education. While in English, such information is coded by a noun, in Brazilian Portuguese, it is a verb that has this function, although both *interest.n* and *interessar-se.v* were created as LUs evoking the Mental_stimulus_experiencer_focus frame. BFN 1.7 defines this frame as follows: "An Experiencer has an emotion as caused by a Stimulus or concerning a Topic".

In the case of English, *interest.n* was created with a non-BFF status in this frame because this noun should actually be part of the MWE *vested interest.n*, which would then have to be created in a frame that contains the entailment that the interest in something is triggered by the fact that such something is of major importance for the collectivity.

Such an entailment is completely lost in the translation of this sentence into Brazilian Portuguese, for which the Mental_stimulus_experiencer_focus frame fits the verb *interessar-se.v* nicely.

Frame	**En LU**	**Br-Pt LU**
Size	huge.a	---
Menta_stimulus_exp_focus	interest.n	interessar-se.v
Degree	---	tanto.adv
Degree	partly.adv	em parte.adv
Causation	because.c	porque.c
Education_teaching	education.n	educação.n
Purpose	mean.v	---
Performers_and_roles	---	papel.n
Bringing	take.v	conduzir.v
Goal	into.prep	---
Temporal_collocation	future.n	futuro.n
Certainty	---	misterioso.a
Capability	can.v	---
Grasp	grasp.v	---

Table 6: Frames for which sentences (5-6) were annotated and LUs evoking them in each language.

However, the difference in the POS of the two LUs has a cascade effect in the other LUs in the sentence. The adverb *tanto.adv 'so much'*, in this sentence, modifies *interessar-se.v 'to be interested in'*. It was annotated in the Degree frame, since it indicates to what degree the speaker is interested in education. Note that the definition of the Degree frame states that "LUs in this frame modify a gradable attribute and describe intensities at the extreme positions on a scale", and, in BFN, the Gradable_attribute FE in this frame is always instantiated as an adjective; there is, however, no reason why gradable attributes cannot be expressed by nouns or verbs. On the other hand, the end-of-scale reading of *tanto.adv*, could not, in this context, be expressed by an adjective such as *huge.a*, which was annotated for the Size frame in the English sentence, generating a frame mismatch in the sentence pair.

This difference, however, does not entail some translation loss, since size and degree are metaphorically linked. On the contrary, they highlight the importance of the net-like configuration of FrameNet at the conceptual - and not only word - level (Fillmore, Baker & Sato 2004) for cross-lingual comparison. In other words, although no obvious word-to-word relation could link and adjective like *huge.a* in English to the adverb *tanto.adv* in Brazilian Portuguese, a metaphor relation connecting the Size and Degree frames could do so.

The same kind of phenomenon is seen in the mismatch between the frames evoked by the predicates indicating the purpose/role of education to take people into the future. Finally, the other differences derive from an inversion, in the translation, of the perspective adopted when talking about the future. In the original English version, the speaker uses a relative clause to modify *future.n*, framing it as something that people do not have the capability to understand. In the translation, the adjective *misterioso 'misterious'* is used to modify *futuro.n*, leaving people's cognitive capacity aside. Even so, the Grasp and the Certainty frame are connected to each other in BFN 1.7 via the Awareness frame. Grasp inherits Awareness, while Certainty uses it. Once again, the fact that FrameNet is a net at the conceptual level sheds some light on differences found in the annotation of a given sentence and its translation.

4. Conclusion

The shared annotation task so far has shown that the frames of Berkeley FrameNet data release 1.7 are complete enough to serve as a basis for cross-linguistic annotation. The initial efforts at annotating an English TED talk and its Brazilian Portuguese translation show that about half of the Brazilian Portuguese frame instances are identical to the frames in English, and about 80% of the Brazilian Portuguese Lexical Units fit without caveat into the frames of BFN 1.7. It also demonstrates that some of the frame mismatches can be better understood if one considers the conceptual-level network of FrameNet. Further research is needed into whether frame definitions based on lexicographic practices are adequate for these kinds of frame mismatches in translations and structural differences between LUs across languages.

5. Acknowledgements

This material is based in part upon work supported by the U.S. National Science Foundation under Grant No. (1629989). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation. The annotation of the TED Talk for Brazilian Portuguese, as well as the development of the FN-Br WebTool were made possible by the CAPES/STINT Grant No. 99999.009910/2014-00 and by the CNPq Grant No. 448990-2014/8.

6. Bibliographical References

Bond, F. and Foster, R. (2013). Linking and Extending an Open Multilingual Wordnet. In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pages 1352–1362, Sofia, Bulgaria, august. Association for Computational Linguistics (ACL).

Fillmore, C. J. (1976). The need for a frame semantics in linguistics. In Karlgren, H. (Ed.) *Statistical Methods in Linguistics*. Stockholm: Scriptor, pp. 5--29.

Fillmore, C. J. (1982). Frame semantics. In The Linguistic Society of Korea (Ed.), *Linguistics in the Morning Calm*. Seoul: Hanshin Publishing Co., pp. 111--137.

Fillmore, C. J., Baker, C. F. and Sato, H. (2004). FrameNet as a "Net". In Maria Tereza Lino (Conference Chair), et al., editors, Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04), pages 1091–1094, Lisbon, Portugal, may. European Language Resource Association (ELRA).

Fillmore, C. J., and Baker, C. F. (2010). A Frames Approach to Semantic Analysis. In B. Heine and H. Narrog, (Eds.), *Oxford Handbook of Linguistic Analysis*. Oxford: Oxford University Press, pp. 313--341.

Hermann, K. M. and Blunsom, P. (2014). Multilingual Models for Compositional Distributed Semantics. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pages 58–68, Baltimore, USA, june. Association for Computational Linguistics (ACL).

Matos, E. and Torrent, T. (2016). A Flexible Tool for an Enriched FrameNet. In 9th International Conference on Construction Grammar (ICCG9), Juiz de Fora, Brazil, october. Federal University of Juiz de Fora (UFJF).

Navigli, R. and Ponzetto, S. P. (2012). BabelNet: The Automatic Construction, Evaluation and Application of a Wide-Coverage Multilingual Semantic Network. *Artificial Intelligence*, 193: 217–250.

Robinson, K. (2006). *Do Schools Kill Creativity?* TED Talk. Video (19 minutes) is available at https://www.ted.com/talks/ken_robinson_says_schools_kill_creativity.

Towards Hindi/Urdu FrameNets via the Multilingual FrameNet

Shafqat Mumtaz Virk[1] **and K.V.S. Prasad**[2]
[1]Språkbanken, Department of Swedish, University of Gothenburg, Sweden
[2]Department of Computer Science and Engineering, Chalmers University of Technology, Sweden
shafqat.virk@svenska.gu.se, prasad@chalmers.se

Abstract

The Multilingual FrameNet Project (MLFN, 2017) is using translations of Ken Robinson's popular TED talk (Robinson, 2006) to study universal and cross lingual aspects of frame annotation. There are no FrameNets yet for Hindi and Urdu, but we are annotating the Hindi and Urdu translations of Robinson's talk using the frames of the English FrameNet. (Surprisingly, there was no Hindi translation, so we did that ourselves). Preprocessing is needed: the word-segmentation and POS tagging tools available for Hindi and Urdu were satisfactory, the full-form lexicons less so. The web-based multi-layer frame annotation tool allows additions to the lexicon, so we simply added each form as a new "word", our goal here being only to look at the frames and frame elements—we plan to look at grammatical function and phrase type later. While some sentences show that the frame analysis of English or Portuguese will not carry over to Hindi or Urdu for cultural or linguistic reasons, others are harder to be definite about. Partly, this is because there are so many possible translations. An expected observation is that a choice of word can steer the focus from one frame to another. Our annotations will help when we start building framenets for Hindi and Urdu.

Keywords: Frame semantics, FrameNet, Multilingual FrameNet, Lexico-Semantic Resources

1. Background: Frame Semantics

Frame semantics, developed by Charles Fillmore and others (Fillmore, 1976; Fillmore, 1977; Fillmore, 1982), thinks of language as creating scenes, in which we understand what a word or phrase means by the role it plays in the scene. E.g., using frame semantics we model a kidnapping *situation* as a structure called a *frame*, a script-like description in which *frame elements* (FEs) such as `Perpetrator, Victim, Purpose, Time` and `Place` play their various roles. Words like `kidnap, abduct, nab` and `snatch` *trigger* this frame. Frames similarly model events, objects, and relations.

Based on frame semantics, a lexico-semantic *FrameNet* (Baker et al., 1998) has been developed since 1998, for English. Descriptions of real world situations are stored as frame scripts in FrameNet, along with the frame elements and triggers that evoke the frame. Each frame is given example sentences, actually occurring text, and there is also a *frame annotated* corpus. The frames are linked by relations to make a FrameNet (henceforth FN). E.g., the frame `Invading` *inherits* from `Attack`, is a *subframe* of `Invasion_scenario`, and *precedes* `Conquering` and `Repel`.

These resources (the FrameNet, the example sentences, and the annotated corpus) have been used for automatic shallow semantic parsing (Gildea and Jurafsky, 2002), itself used in 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).

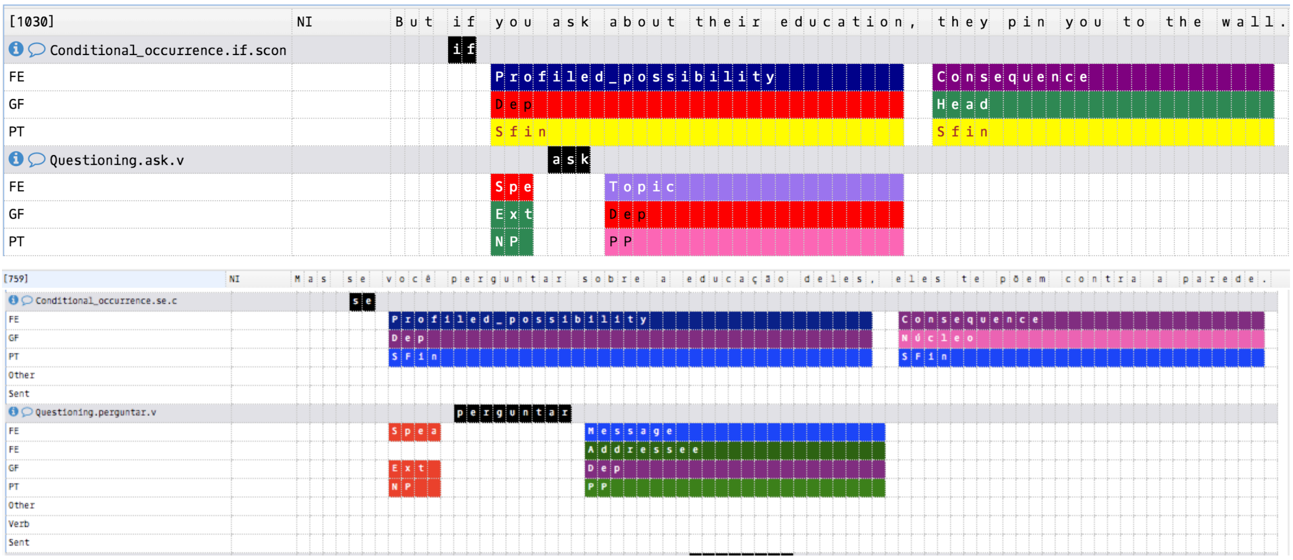
2. MultiLingual FrameNet

FrameNets have since been built for several languages (Chinese, French, German, Hebrew, Korean, Italian, Japanese, Portuguese, Spanish, and Swedish), and have helped explore various semantic characteristics of the individual languages, but the cross linguistic and universal aspects of the FN model are largely yet to be studied. So a MultiLingual FN (MLFN) is now being built by aligning FNs of the individual languages. As a first step, translations of Ken Robinson's popular TED talk (Robinson, 2006) are being annotated using the frames of the Berkeley English FrameNet. An example annotation is shown in Fig. 1

Annotators for each language mark the frame-elements (FE), the grammatical function (GF), and the phrase type (PT) of the marked FEs. (See Sec. 5. for a brief description of these layers, and (Ruppenhofer et al., 2006) for more details). Fig. 1 shows the annotations of two frames, `Conditional_occurrence` and `Questioning`, in the sentence "But if you ask about their education, they pin you to the wall". These are triggered respectively by the lexical units `if` and `ask`. The text blocks "you" and "about their education" have been marked as FEs `Speaker` and `Topic` respectively. The GF and PT of the marked FEs have been labeled at their corresponding layers. (Ruppenhofer et al., 2006) explains the PT and GF labels for English.

Annotators choose from a given list of frames and their FEs. If an annotator does not find a suitable frame from the given list, they select the best alternative (if any), note why the frame is unsuitable, and suggest a better frame. The PT and GF are language dependent, and a list of PTs and GFs for each language has to be provided by the annotators.

Fig. 1 also shows the Portuguese translation of the sentence with the same two frames. Note that a different FE, `Message`, used for asking "What is this", is chosen instead of `Topic`, used for "asked about train times". Whether the choice is appropriate is up to the



[1030] NI But if you ask about their education, they pin you to the wall.

Conditional_occurrence.if.scon — if
FE: Profiled_possibility | Consequence
GF: Dep | Head
PT: Sfin | Sfin

Questioning.ask.v — ask
FE: Spe | Topic
GF: Ext | Dep
PT: NP | PP

[759] NI Mas se você perguntar sobre a educação deles, eles te põem contra a parede.

Conditional_occurrence.se.c — se
FE: Profiled_possibility | Consequence
GF: Dep | Núcleo
PT: SFin | SFin
Other
Sent

Questioning.perguntar.v — perguntar
FE: Spea | Message
FE: Addressee
GF: Ext | Dep
PT: NP | PP
Other
Verb
Sent

Figure 1: Partial frame-annotation of an example sentence in English and Portuguese.

annotator.

Once this multi-lingual annotation is completed, the challenges faced by annotators, the common and uncommon frames chosen, and the attached notes, will be collated and reported. Similarly, variations in PTs and GFs of the various FEs. These reports will be used to learn about the difficulties and challenges in trying to align existing framenets, and building a multilingual framenet.

This paper reports our experience of annotating the Hindi and Urdu translations of Robinson's talk.

3. Background: Hindi and Urdu

'Hindi'[1] has ca. 400 m (million) speakers, of whom 250 m are native. Urdu[2] has ca. 250 m speakers, of whom 60 m are native. Only English, Mandarin, Spanish and Arabic have more speakers than Hindi-Urdu.

Hindi and Urdu 'share the same grammar and most of the basic vocabulary of everyday speech', but are 'two separate languages in terms of script, higher vocabulary, and cultural ambiance' (Flagship, 2012; Prasad and Virk, 2012). They are thus different standard registers of one language (Bhat et al., 2016). Indeed, we used a tool (Apertium, 2017) that translates efficiently between the two, doing mostly only lexical substitution.

Hindustani. The 'Hindi' of films and songs is 'the common spoken variety, devoid of heavy borrowings from either Sanskrit or Perso-Arabic' (Kachru, 2006). We call this form *Hindustani* (Chand, 1944; Bailey et al., 1950). India's 'Hindi' belt speaks more Hindustani than Hindi. But Hindustani has no 'status in Indian or Pakistani society' (Kachru, 2006). We study only Urdu and (standard) Hindi[3] here.

Scripts. Hindustani[4] began ca. 1400 as a Delhi dialect with some Perso-Arabic vocabulary. Urdu, ca. 1750, is Hindustani with copious Perso-Arabic borrowings. Both are written in Perso-Arabic script.

By 1900, some began to write Hindustani in Devanagari[5], the script giving it an identity, *Hindi*, distinct from Urdu, and an impetus to progressively use Sanskrit vocabulary instead of Perso-Arabic.

The Hindi lexicon. Hindi and Urdu 'share the same Indic[6] base' (Schmidt, 2004), and a phonology (UH) that breaks up the consonant clusters of Sanskrit, and drops short vowels at the end of syllables.

Phonology plays no role in frame analysis, but that the phonologies of Sanskrit and UH are at odds is a feature of the Hindi lexicon, which does matter.

E.g., suppose we replace the Indic Hindi-Urdu word सूरज `sūraj` "sun" with the Sanskrit सूर्य `sury`. In Sanskrit, सूर्य is pronounced `surya`, easy to say, but UH drops the final `a` in speech, producing a hard-to-say word-final consonant cluster. (Dropped vowels remain in the script, and re-appear as schwas in song).

Other awkward Sanskrit words are e.g. यदि `yadi` "if", परन्तु `parantu` "but", शक्ति `šakti` "power", with their short vowel endings, a feature foreign to UH.

Unadapted Sanskrit words make Hindi more "national", but sound odd. Older Indic literary languages with UH phonology and adapted Sanskrit borrowings are not 'in direct linguistic antecedence to [...] Hindi.

[1] By Hindi, we mean standard Hindi. We take 'Hindi' more broadly, including its many dialects, some being arguably distinct languages. Multiple lexicons give 'Hindi' multiple *forms* (Kachru, 2006).

[2] Urdu has always drawn its advanced vocabulary only from Perso-Arabic, and has basically just one form.

[3] In Hindustani and in the 'Hindi' belt, "sky" is आसमान `āsmān`, a Persian word. In Hindi and other Indian languages, it is आकाश `ākāš`, a Sanskrit word. The preference for Sanskrit makes Hindi better understood nationally.

[4] Also called 'Hindi' then, but we reserve this term for the modern language, to reduce confusion.

[5] The script used for Sanskrit.

[6] i.e., with no Perso-Arabic words.

The one language that is antecedent [is] Urdu ...' (Masica, 1991). The lexical future will be interesting.

4. Translating the Urdu text to Hindi

An Urdu translation of Robinson's talk was available when we started, but surprisingly, not a Hindi one. We produced one ourselves (one of us speaks Hindi, but is not native), starting by pushing the Urdu text through Apertium, which fortunately has an Urdu-Hindi pair implemented. The output included much text that was just a transcription from Perso-Arabic script to Devanagari, as well as some Urdu text where even the transcription failed. These might be seen as shortcomings, but we think they are outweighed by the sensible behaviour of the tool in keeping going—the user will have to edit the output anyway, and these errors are easy to spot.

The manual corrections needed took several days full time, though experience with other languages suggests this is still less time than a translation from scratch would take. Finally, our text was validated and improved by a native Hindi speaker.

5. Pre-processing

We do frame annotations using the MLFN version of the Berkeley English FrameNet webtool. It allows us to attach syntactic and semantic annotation layers to the subject text. To set up the tool for a given language, the following data files are needed. Given the size of Hindi-Urdu, it is odd that we sometimes didn't find the needed resources. Those working with other South Asian (SA) languages may face similar situations.

1. A sentence segmented UTF text. We could find no publicly available sentence segmentors for either Hindi or Urdu, so we used a program to split the text at particular punctuation symbols, and then validated the results by hand.
2. A file listing all word forms of all the lexemes in the text together with the part of speech (POS) tag of each lexeme. For this, we used the smart morphological paradigms of GF (Virk et al., 2010). These take a word, and based on word endings and other clues, attempt to find suitable word-formation functions to build inflection tables. However, they are still occasionally error-prone and also have limited coverage. Fortunately, the MLFN tool allows additions to the lexicon, so we simply added each surface form as a new "word" as we went along.
3. We used the universal POS tagger for Hindi to tag the text, and the tags were then mapped to the FrameNet POS tagset [7]. For Urdu POS tagging, we used curlp Urdu POS tagger[8].
4. A list of annotation labels to be used for each language. For this experiment, Frame Element (FE), Phrase Type (PT), and Grammatical Function (GF), layers are to be added. Details can be found in the FrameNet book (Ruppenhofer et al., 2006). We briefly describe the only three annotation layers needed at this stage.

   Frame Element (FE) Here, annotators choose a suitable FE label. E.g., `Topic` in Fig. 1. Labels are taken from FrameNet data release 1.7, and annotators are not allowed to change them.

   Phrase Type (PT) Here, annotators classify the text that makes up each FE. The set of PTs is language dependent, will be chosen by the annotation team. For Hindi and Urdu, we opted to start with the English PTs, and add/edit types as needed (the MLFN tool allows these actions).

   Grammatical Function (GF) Annotators assign a GF to each FE, saying how the FE satisfies its grammatical requirements (Ruppenhofer et al., 2006). The set of GFs too is language dependent, but we opted to start with the English GF labels.

6. Annotation Status

Table 1 shows statistics of the annotations done so far both for Hindi and Urdu. For Hindi, a total of 84 frames and 154 frame-elements were annotated from the first 25 sentences of the talk. As can be noted, most of the lexical units (i.e. triggers) are from the noun and verb class followed by adjectives and adverbs. The remaining lexical units are conjunctions, prepositions and numbers. For Urdu, a total of 42 frames and 76 frame-elements were annotated from the first 27 sentences of the talk.

	Hindi	Urdu
Sentence	25	27
Frames	84	42
Frame-Elements	154	76
Noun Triggers	25	17
Verb Triggers	22	16
Adjective Triggers	13	6
Num Triggers	3	2
Adverb Triggers	8	1
Prep Triggers	3	-
Conjunction Triggers	10	-

Table 1: Annotation Statistics

7. Observations and Lessons

Some example sentences from Robinson's talk, where cross-lingual annotation is expectedly problematic: idiom ("good morning"), slang ("I've been blown away"), and metaphor ("themes running through").

1. Good morning.

   In Hindi, this is नमस्ते `namaste` "Greetings". There are no separate greetings for times of day, or even

[7] FN tagset: 'A' = Adjective, 'ADV' = Adverb, 'ART' = Article, 'AVP' = Adverbial Preposition, 'C' = Conjunction, 'INTJ' = Interjection, 'N' = Noun, 'NUM' = Number, 'PREP' = Preposition, 'PRON' = Pronoun, 'V' = Verb.

[8] For a demo,see http://182.180.102.251:8080/tag

to say "hello" or "bye". The occasion may be marked by other sentences.

In Urdu,

صبح بخیر

subah buxair "Good morning"

2. I've been blown away by the whole thing.

   In Hindi, this is मेरी तो बुद्धि ही उड़ गयी है
   merī to buddhi hī uṛ gayī hai
   "my mind itself has been blown away".

   In Urdu,

   مجھے تو اس سب نے ہلا کر رکھ دیا ہے

   mujhe to is sab ne hilā kar rakh diyā hai
   "As for me, all this has left me shaken".

   A slang expression, this is hard to translate. In both English and Hindi, the verb blown evokes the frame Motion, but the FE Theme changes from me to my mind. Urdu changes the frame to Cause_to_move_in_place, but the FE Theme is again me.

3. There have been three themes running through the conference.

   In Hindi, सम्मेलन में तीन विषय उभर कर आ रहे हैं
   sammelan mẽ tīn viṣay ubhar kar ā rahe hãı
   "in the conference, three things are coming up".

   The English running evokes the frame Fludic_motion, with FEs Fluid "three themes" and Area "through the conference". The Hindi ubhar kar ā evokes Coming_to_be with FEs Place "in the conference", Entity "three things" and Time "are ...ing". Both are idiomatic expressions, and a different Hindi translation might have used the image of three streams flowing.

Most of the few dozen sentences we have annotated so far pose more interesting questions since the differences are not as easily explained away as in the above examples. Unfortunately, these few dozen are not enough to observe patterns in bulk. For when we have a larger number, we anticipate a few features and challenges.

Causation Where the intransitive verb "shake" evokes Motion, the transitive verb evokes Cause_to_move_in_place, as in example 2. In Hindi-Urdu this shift is done morphologically, by making causative verbs out of intransitive ones. Thus hilnā "to shake (intr.)" becomes the hilānā "to shake (tr.)" of example 2.

Examples abound: khānā "to eat" and khilānā "to feed", sonā "to sleep" and sulānā "to put to bed", etc., where English uses a different verb or an auxiliary causative verb.

Hindi-Urdu also have verbs for indirect causation. hilvānā, khilvānā, sulvānā mean to get somebody else to shake (tr.), feed, and put to bed. Even when the basic verb is transitive, such as "sell" becnā with its causal version bicvānā "get sb. to sell", there may be a kind of back-formation to the intransitive verb: biknā, used to say something sells well/badly, or is available for sale.

These regular causative links can perhaps be reflected in FN; of interest because this feature appears in other SA languages.

Abstract or concrete? A sentence in the talk is "Because it's one of those things that goes deep with people", where deep evoked frame Measurable_attributes. The Urdu text maintained the abstraction: "it goes into the depth in people". Our Hindi informant preferred "it lives in the depths of the heart(s) of people", more concrete and evoking the frame Body_parts.

A similar example is "a future that we can't grasp", where the verb evoked Grasp. Again, our Hindi text is more concrete: "that is outside our imagination", evoking Image_schema.

It is unlikely that such cases will show a systematic variation in frame choice going from English to Hindi-Urdu, beyond suggesting many new frames (heart, imagination, etc.).

Verb or noun? The Urdu text for "future we can't grasp" is "doesn't come into our grasp", a verb-noun variation that may be systematic. The frames evoked are different, but the meaning is the same, suggesting we look for higher level frames. Hindi-Urdu has a range of nouns that come from verbs, and vice-versa, as does English. Frame connections even within Hindi-Urdu may be interesting, as with causation.

Complex lexemes "Come into grasp" can be seen as a *complex lexeme*, a verb-based multi-word expression (Hook, 1974; Masica, 2005). In the English FN lexicon, there are many lexical units which will correspond to such complex lexemes in Hindi/Urdu. The status of these constructions as lexical or grammatical is debated and they are generally under-researched (Schultze-Berndt, 2006; Butt, 2010; Slade, 2016)

8. Conclusions and Outlook

We have started annotating Hindi/Urdu using the MLFN tool, and have reported on our experience so far. We are some way from being able to note systematic changes in annotation going from, say, English to Hindi or Urdu, and we have even further to go to construct FrameNets for Indic languages. But we can already say confidently that despite the shortage of resources, our exercise has been worthwhile and we would encourage similar work on other SA languages. Two lessons to note:

First, translation and frame annotation teach us much about the target languages.

Second, provided the target language has at least rudimentary dictionaries and enough text online to help the novice writer, a translator can start with not much more than an ability to speak the language. They can learn as they go. Indeed, the TED translations are crowd-sourced. This is one way to rapidly add publicly available texts, a big help for poorly resourced languages. The quality will be variable, but can be improved afterhand. Meanwhile, the crowd-sourcing builds up an even more valuable resource: a community with greater competence in the target languages.

Annotation needs access to the tool, and some training, but not too much. Here too, one might be able to use volunteers to help, thus building up a FrameNet, and a full form lexicon.

For future work, we list some features of Hindi-Urdu, many shared with all SA languages, both Indo-European and Dravidian. We want to know how these features affect frame analysis. The data we gather will help us build FNs for Hindi and Urdu.

Reduplication is a prominent feature of all SA languages. It can mean greater intensity, or longer duration, but also distribution: "give the children two-two pencils" means "give each child two pencils".

States of mind. In SA languages, "I am hungry" and "I like spinach" are both expressed "to me, hunger affects" and "to me, spinach liking affects". Note that the English verbs can be transitive or intransitive. Is there a regular change in frames and triggers?

Clitics Examples are `hī` and `to` in Example 2 of Section 6, translated crudely as "itself" and "as for". These are function words, and it is hard to see of hand how they might affect choice of frame, but they do change the meaning of a sentence.

PTs and GFs. We have yet to work these out.

Incompatible lexicons. Hindi and Urdu differ only in lexicons, but the words are not one-to-one equivalents, at best they overlap largely. Some social or religious structures don't map at all, and the words have to be borrowed. An apparently equivalent word might require a different grammatical structure. So we expect the FrameNets to be affected by these factors, and cast light on them.

Cultural factors. Translating from English to SA languages involves a huge change of culture, and we can expect interesting new frames and compromises in translations. E.g., the Hindi "good morning" in Sec. 6. For more spectacle, consider weddings in the various Western and Asian communities.

9. Acknowledgements

We thank the anonymous referees for their detailed comments, which we have incorporated into the text above.

The work presented here has been financially supported by the University of Gothenburg, its Faculty of Arts and its Department of Swedish, through their truly long-term support of the Språkbanken research infrastructure, 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), *Swe-Clarin* (2014–2018; contract 821-2013-2003), and the Department of Computer Science and Engineering, Chalmers University of Technology, Sweden.

References

Apertium. (2017). Apertium Wiki main page. `http://wiki.apertium.org/wiki/Main_Page`.

Bailey, T. G., Firth, J. R., and Harley, A. H. (1950). *Teach yourself Hindustani.* English Universities Press, London. Available from archive.org.

Baker, C. F., Fillmore, C. J., and Lowe, J. B. (1998). The Berkeley FrameNet Project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*, ACL '98, pages 86–90, Stroudsburg, PA, USA. Association for Computational Linguistics.

Bhat, R. A., Bhat, I. A., Jain, N., and Sharma, D. M. (2016). A House United: Bridging the Script and Lexical Barrier between Hindi and Urdu. In *COLING*, pages 397–408. ACL.

Butt, M. (2010). The light verb jungle: Still hacking away. In Mengistu Amberber, et al., editors, *Complex predicates: Cross-linguistic perspectives on event structure*, page 48–78. Cambridge University Press, Cambridge.

Chand, T., (1944). *The problem of Hindustani. Allahabad: Indian Periodicals.* `www.columbia.edu/itc/mealac/pritchett/00fwp/sitemap.html`.

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. In Antonio Zampolli, editor, *Linguistic Structures Processing*, number 59 in Fundamental Studies in Computer Science. North Holland Publishing.

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

Flagship. (2012). *Undergraduate program and resource center for Hindi-Urdu.* University of Texas at Austin. `http://hindiurduflagship.org/about/two-languages-or-one/`.

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

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.

Hook, P. (1974). *The Compound Verb in Hindi.* Michigan series in South and Southeast Asian languages and linguistics. University of Michigan, Ann Arbor.

Kachru, Y. (2006). *Hindi (London Oriental and African Language Library).* Philadelphia: John Benjamins Publ. Co.

Liu, D. and Gildea, D. (2010). Semantic role features for machine translation. In *Proceedings of COLING 2010*, COLING '10, pages 716–724, Beijing. ACL.

Masica, C. (1991). *The Indo-Aryan languages.* Cambridge University Press.

Masica, C. (2005). *Defining a Linguistic Area: South Asia.* Chronicle Books.

MLFN. (2017). Multilingual FrameNet Project. framenet.icsi.berkeley.edu.

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

Prasad, K. V. S. and Virk, S. (2012). Computational evidence that Hindi and Urdu share a grammar but not the lexicon. In *3rd Workshop on South and Southeast Asian Natural Language Processing (SANLP), collocated with COLING 12.*

Robinson, K. (2006). Do schools kill creativity? *TED: Ideas worth spreading.* https://www.ted.com/talks/ken_robinson_says_schools_kill_creativity/up-next.

Ruppenhofer, J., Ellsworth, M., Petruck, M. R., Johnson, C. R., and Scheffczyk, J. (2006). *FrameNet II: Extended Theory and Practice.* International Computer Science Institute, Berkeley, California. Distributed with the FrameNet data.

Schmidt, R. L. (2004). *Urdu: An Essential Grammar.* London/ New York: Routledge. See the preface by Gopi Chand Narang.

Schultze-Berndt, E. (2006). Taking a closer look at function verbs: Lexicon, grammar, or both? In Felix K. Ameka, et al., editors, *Catching language: The standing challenge of grammar writing*, page 359–391. Mouton de Gruyter, Berlin.

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

Slade, B. (2016). Compound verbs in Indo-Aryan. In Hans Henrich Hock et al., editors, *The languages and linguistics of South Asia: A comprehensive guide*, pages 559–567. De Gruyter Mouton.

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

Virk, S. M., Humayoun, M., and Ranta, A. (2010). An open source Urdu resource grammar. In *Proceedings of the Eighth Workshop on Asian Language Resouces*, pages 153–160, Beijing, China, August. Coling 2010 Organizing Committee.

Wu, D. and Fung, P. (2009). Semantic roles for SMT: A hybrid two-pass model. In *Proceedings of HLT-NAACL 2009*, NAACL-Short '09, pages 13–16, Boulder. ACL.

Towards an Open Dutch FrameNet lexicon and corpus

Piek Vossen, Antske Fokkens, Isa Maks, Chantal van Son
Vrije Universiteit Amsterdam
De Boelelaan 1105, 1081 HV Amsterdam
The Netherlands
{piek.vossen, antske.fokkens, isa.maks, c.m.van.son}@vu.nl

Abstract

This paper reports on the progress of the development of an Open Dutch FrameNet lexicon and annotated corpus. We started the project in 2017 with the annotation of a Dutch corpus of written Dutch that was previously annotated with PropBank predicates and roles. The corpus represents a diverse set of written Dutch texts. We discuss the annotation results and process. From this corpus, we have derived an initial Dutch lexicon with FrameNet frames. In the meanwhile, we designed a method to collect texts that exhibit a large degree of variation in framing similar events. We will apply this method in the future to extend the representative corpus vertically for certain types of events to obtain more insight into variation of framing.

Keywords: Dutch, frame semantics, corpus annotation

1. Introduction

Languages are rich instruments for framing situations or events in various ways. A report on a football game, for instance, can be written from the perspective of the winner, the loser, or a neutral observer; a financial transaction can be reported from the buyer or the seller; a medical case can be framed from the perspective of the patient or the doctor. We use different words and expressions in language to frame similar situations differently depending on our interest, our motivation, and audience. The perspective on a situation that is associated with the choice of words is what we call linguistic framing. It reflects what we see as important and what as background, it expresses emotions and judgments, and it suggests motivations and expectations. A concrete case in point is work by Cybulska and Vossen (2010), who demonstrate how the *Fall of Srebrenica* is framed differently depending on the time passed between the event taking place and the moment of reporting. As historic distance increases, less detail (e.g. abstracting from the precise time, location and participants) but more explanations, motivations and judgments (deportation, genocide) were given. Fokkens et al. (2018) investigate how stereotypes and created images are reflected in textual micro-portraits (framings of individuals in stories) and show, for instance, that Dutch newspapers mostly specifically label people as "Dutch" when they win in sports.

Clearly, language is a powerful instrument to shape our view of the world, and it is therefore important to get a good understanding of how framing works. Yet, little is known about framing in Dutch. What are the Dutch words and expressions used to frame the same situations or events in different ways? How does Dutch framing differ from other languages? How much variation exists and what are the underlying semantic and pragmatic factors for using these variants in contexts?

This paper reports on the initial development of the Open Dutch FrameNet similar to multilingual FrameNets described in (Baker, 2008). We started the development of a Dutch FrameNet in 2017 with the annotation of a corpus of written Dutch that was previously annotated with PropBank predicates and roles (Kingsbury and Palmer, 2002); see Sections 2. and 3.. From this corpus, we derived an initial FrameNet lexicon (Section 4.). For future work (Section 5.), we will use a method to collect texts that exhibit a large degree of variation in framing similar events.

2. Overall Approach

Our first objective is to capture the usage of FrameNet frames and elements in a representative Dutch corpus and to derive a Dutch FrameNet lexicon from this corpus. We therefore took the following design decisions:

- We use a balanced corpus with diverse genres;
- We apply an all-sentences-approach, which means:
  - we take the sentences of a document as given
  - we do not apply any preselection of lexical units nor a preselection of example sentences;
  - we also do not preselect frames or frame elements;
  - but for each sentence a preselection of the main predicate and the arguments is already given;
- Frame identification should fit the usage of the predicate in the sentence;
- Roles are assigned after the sentence-frame is selected with the corresponding roles.

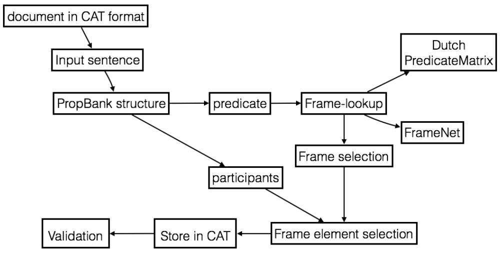
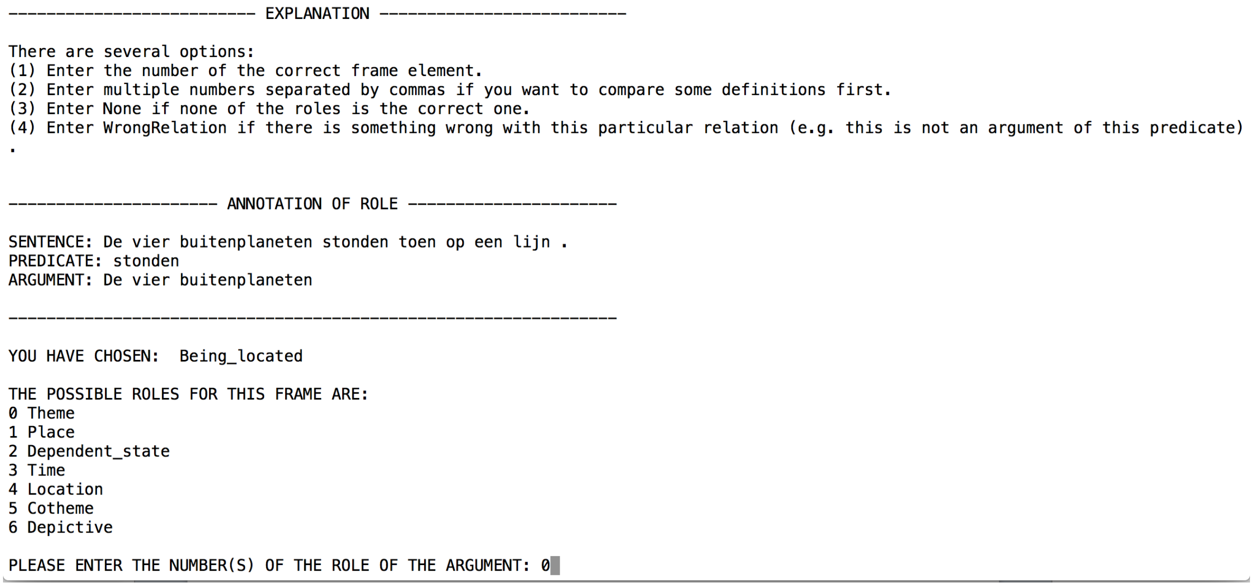


Figure 1: Overview of the annotation process of the SoNaR documents with PropBank annotation in the CAT format.



```
------------------------- EXPLANATION -------------------------

There are several options:
(1) Enter the number of the correct frame element.
(2) Enter multiple numbers separated by commas if you want to compare some definitions first.
(3) Enter None if none of the roles is the correct one.
(4) Enter WrongRelation if there is something wrong with this particular relation (e.g. this is not an argument of this predicate)
.

--------------------- ANNOTATION OF ROLE ----------------------

SENTENCE: De vier buitenplaneten stonden toen op een lijn .
PREDICATE: stonden
ARGUMENT: De vier buitenplaneten

---------------------------------------------------------------

YOU HAVE CHOSEN:  Being_located

THE POSSIBLE ROLES FOR THIS FRAME ARE:
0 Theme
1 Place
2 Dependent_state
3 Time
4 Location
5 Cotheme
6 Depictive

PLEASE ENTER THE NUMBER(S) OF THE ROLE OF THE ARGUMENT: 0
```

Figure 2: Screenshot of the annotation interface showing instructions, the target sentence and the target predicate and an argument according to the PropBank structure for which a frame element needs to be selected, given the frame Being_located that was assigned to the predicate *stonden* (stood).

We used SoNaR as a corpus, which is a corpus of written Dutch (Oostdijk et al., 2008). Part of this corpus was already annotated with PropBank relations (De Clercq et al., 2012). Figure 1 shows the further process starting with documents from SoNaR in the format of the CAT annotation tool (Lenzi et al., 2012). Our annotators first add FrameNet annotations to these previously annotated PropBank predicates (verbs) and their arguments. Because the annotators proceed sentence-by-sentence through a highly varied set of texts, they have to consider all frames from the English FrameNet version 1.7. We therefore developed a specific annotation tool[1] to support the annotators, which loads the annotated PropBank relations one by one and presents the annotator with the sentence, the predicate and the arguments. The annotation task consists of two steps: (1) frame annotation, and (2) frame element annotation. For the first step, the tool supports searching for frames in FrameNet by entering the predicate and/or equivalents in both Dutch and English. Equivalents are generated using the PredicateMatrix (derived from SemLink (De Lacalle et al., 2014)), which provides mappings between English and Dutch lexical units through the Open Dutch WordNet (Postma et al., 2016). After entering the predicate and/or equivalents, the annotator is then presented with the definitions of all associated frames and selects the most fitting one (if any). More experienced annotators can also directly enter the name of the frame. Once a frame is selected for the predicate, the tool iterates over the arguments to select the frame elements. Figure 2 shows a screenshot of the frame element annotation after the frame Being_located has been selected for the sentence in Example 1 from the Dutch Wikipedia article on the solar system.

(1) *De vier buitenplaneten* ***stonden*** *toen op een lijn*
The four outer planets **stood** then in one line.
"The four outer planets were aligned in those days."

Texts annotated by two annotators are processed to mark mismatches and disagreement. We distinguish between mismatches between frames that stand in a super-subtype relation in FrameNet and other mismatches. Texts with marked agreement and disagreement are visualised for analysis and adjudication using the CAT tool.

3. Frame Corpus

Four students worked for four months, eight hours a week. All texts have been double annotated. In total, 3,898 verb tokens have been annotated with 679 frames. Table 1 shows the statistics for the annotated corpus, showing the distribution of texts and the number of annotated predicates for each genre. The most represented genres are financial, periodicals and wikipedia.

theme/genre	nr_of_files	nr_of_annotated_verbs
background-news	3	110
financial	17	1756
medical	1	88
news	5	499
newsletter	3	111
periodicals	37	821
policy	12	352
teletext	3	169
websites	1	49
wiki	34	1295
totals	116	5250

Table 1: Corpus statistics on the different genres and the number of files in the SoNaR corpus that have PropBank annotations with the total number of annotated predicates in each genre.

[1] https://github.com/cltl/FrameNet-annotation-tool

We measure the inter-annotator agreement (see Table 2) counting exact matches (47%, Kappa 0.46) and lenient matches. In the case of lenient matches, we consider frames to be matches if they are closely related by one of FrameNet's frame-to-frame relations such as Inheritance (lenient agreement-I) or any relation (lenient agreement-II). Inter-annotator agreement increase with 3% and 7% respectively when lenient matching is applied. Agreement in annotating frame elements given agreement on the frame was much higher (79%). Frame agreement is lower than agreement scores reported by, for example, Søgaard et al. (2015) and Benešová et al. (2008), who respectively report scores of 85% (frames) and 78% (frame elements) on English Twitter data, and 69% and 85% on Czech lexical units for communication verbs. However, in these studies, the annotation tasks were much more restricted in the types and/or number of frames to be considered. Following the procedure explained in the previous section, our annotators need to proceed sentence-by-sentence, considering very different predicates and all types of frames and all possible relations.

Type of agreement	Percentage
strict agreement	0.47
lenient agreement -I: only inheritance relations	0.51
lenient agreement -II: all relations	0.54
agreement on frame elements (with matching frames)	0.79

Table 2: Inter-annotator agreement statistics on frames and frame elements.

The annotators struggle both with consistently selecting frames from the large set available in FrameNet and with coverage problems of FrameNet (in which case the frame "None" is assigned). In Table 4, we show the most frequently confused frames. As was also found by Padó (2007, p. 63), some of these disagreements are due to subtle or difficult distinctions between frames in meaning that may not be clear from the context. Therefore, we further analyzed the disagreements by determining the distance between the confused frames in the frame hierarchy (taking all relations into account) and the type of relations between them. We found that in 20% (552 instances) of all disagreements, the frames were directly related through one of the ten frame relation types in FrameNet (frame-frame distance of 1). The distribution of the relation types in these cases is shown in Table 3. For example, there is an Inheritance relation between many of the most frequent frame confusion pairs, e.g. {Activity_start, Process_start}, {Creating, Intentionally_create}. Other frequent cases include those frames standing in a Using relation; for example, the frame Communication is used in many other frames, such as Statement and Expressing_publicly. The ReFraming_Mapping relation between two frames indicates that lexical units were moved into a new frame (Petruck et al., 2004), as is the case for the pair {Attempt_suasion, Request}. In many of these cases, one frame may be more specific than the other, but both are likely to fit the lexical unit found in the text. For example, both Creating and Intentionally_create are technically correct for the lexical unit *maken* in Sentence 2, even though Intentionally_create would be more specific.

(2) *maar wij moeten het beter doen en minder van die regels* ***maken***
but we must it better do and less of those rules **make**
"but we have to do better and make less of those rules."

Frame relation type	Percentage
Inheritance	0.40
Using	0.21
ReFraming_Mapping	0.14
Causative_of	0.12
See_also	0.09
Inchoative_of	0.01
Perspective_on	0.01
Precedes	0.01
Subframe	0.01
Metaphor	0.0

Table 3: Distribution of types of relations between confused frames with a frame-frame distance of 1.

Other confusions, however, seem to involve frames with different core elements and restrictions on these core elements (such as +CONTROL or -CONTROL) which are not likely to be both correct for one context, as with the pair {Operate_vehicle, Self_motion}. However, even these distinctions are not always clear. For example, the correct frame in Sentence 3 for *gereden* seems to be Operate_vehicle, whereas Self_motion seems less correct. However, the definition of Self_motion does mention that "many of the lexical units in this frame can also describe the motion of vehicles (e.g., as external arguments) [and are treated] as belonging in this frame."

(3) *Doorgaans wordt vanwege de risico's in konvooi* ***gereden***
Usually being because the risks in convoy **driven**
'Usually, vehicles are **driven** in convoy because of the risks."

The other frame confusion pairs had a frame-frame distance of two (15%), three (17%), more (42%), or were not related at all (7%). Even though frame confusions were never counted as correct in our agreement scores if their frame-frame distance is larger than one, some of them are still understandable. For example, the frames Daring and Attempt are not directly related to each other, but both inherit from Intentionally_act, which makes them sister frames (distance=2). We also encountered 'grandparent' relations, such as {Finish_competition, Activity_finish} linked through Finish_game (distance=2). Frame pairs with larger distances are more likely to exhibit significant semantic differences, as with {Path_shape, Sign_agreement} (distance=5), but not necessarily, as with {Opinion, Regard} (distance=5).

In Table 5, we show agreement and disagreement for the most frequent frames. We can see that the (dis)agreement varies considerably across frames: e.g. Desiring (69), Attempt_suasion (65) and Statement (64) as highest scoring and Circumscribed_existence (6), Intentionally_create (7) as lowest scoring. High agreements could be due to frequency of certain predicates with clear meaning and little ambiguity. Low agreements seem idiosyncratic.

Our annotations are open source and freely downloadable as well as some of the original texts.[2] Part of the original

[2] https://github.com/cltl/Open-Dutch-Framenet

19	Activity_start	Process_start
14	Creating	Intentionally_create
14	Cause_change_of_position_on_a_scale	Change_position_on_a_scale
12	Using	Using_resource
12	Opinion	Regard
10	Cooking_creation	Manufacturing
8	Getting	Receiving
8	Expressing_publicly	Statement
8	Existence	Presence
8	Awareness	Grasp
7	Operate_vehicle	Self_motion
7	Finish_competition	Finish_game
7	Causation	Evidence
7	Being_named	Name_conferral
6	Perception_active	Perception_experience
6	Intentionally_create	Text_creation
6	Giving	Grant_permission
6	Cure	Medical_intervention
6	Cause_to_perceive	Expressing_publicly
6	Beat_opponent	Finish_competition
6	Awareness	Certainty
6	Accomplishment	Getting
5	Reference_text	WrongRelation
5	Preventing	Thwarting
5	Perception_active	Reference_text
5	Have_associated	Possession
5	Finish_competition	Success_or_failure
5	Competition	Finish_competition
5	Communication	Statement
5	Communication	Expressing_publicly

Table 4: Frame confusion pairs across annotators sorted by frequency.

frame	agreements	disagreements	percentage agreement
Desiring	25	11	0.69
Attempt_suasion	33	18	0.65
Statement	108	60	0.64
Request	19	19	0.5
Removing	16	18	0.47
Causation	38	45	0.46
Receiving	22	31	0.42
Self_motion	17	23	0.42
Change_position_on_a_scale	27	39	0.41
Perception_active	17	24	0.41
Activity_start	23	35	0.4
Event	30	46	0.39
Coming_to_be	30	46	0.39
Being_located	22	35	0.39
Using	17	28	0.38
Cause_change	13	21	0.38
Possession	38	64	0.37
Intentionally_act	24	49	0.33
Opinion	18	38	0.32
Participation	12	27	0.31
Existence	26	62	0.3
Evidence	14	34	0.29
Becoming_aware	12	33	0.27
Inclusion	12	39	0.24
Process_start	8	26	0.24
Arriving	8	26	0.24
Cause_to_perceive	11	36	0.23
Awareness	14	50	0.22
Accomplishment	6	27	0.18
Giving	7	35	0.17
Communication	6	30	0.17
Finish_competition	5	35	0.12
Intentionally_create	4	52	0.07
Circumscribed_existence	2	32	0.06
None	2	40	0.05

Table 5: Most frequently assigned frames with the agreements and disagreements.

texts must however be obtained through a license (freely available for research): "SoNaR-klein-commercieel" enriched with PropBank annotations.[3]

4. Initial frame lexicon

We can derive an initial FrameNet lexicon for Dutch from the annotations made so far. In total more than 1,336 predicate types or lexical entries have been annotated. We list all the different frames that have been assigned to these predicates with their frequency. If we consider each lemma-frame pair as a lexical unit, we would get 4,755 different lexical units distributed across 671 frames. Figure 3 shows a few examples of this derived lexicon. We see that the annotator assigned six different frames to the polysemous Dutch word *afsluiten* (close, settle, end). Some of these frames are closely related to each other representing three of the main meanings of the word: the meaning *close a building or door* is represented by the frames Locale_closure and Change_activity, the meaning *settle an agreement* is represented by Make_agreement_on_action and Sign_agreement and the meaning *finish a process* by Activity_finish and Process_end. The example shows that in this way not only coarse-grained senses, but also more fine-grained nuances of word senses are captured.

```
<?xml version="1.0" encoding="UTF-8" standalone="yes"?>
<fnLexicon lang="nl">

<ENTRY lemma="inschakelen" pos="v">  /* switch on */
 <frameAnnotation frame="Installing" annotations="1"/>
 <frameAnnotation frame="Process_start" annotations="1"/>
</ENTRY>
<ENTRY lemma="mankeren" pos="v">  /* be inadequate */
 <frameAnnotation frame="Medical_conditions" annotations="2"/>
 <frameAnnotation frame="Undergoing" annotations="2"/>
</ENTRY>
<ENTRY lemma="baseren" pos="v">  /* base on */
 <frameAnnotation frame="Evidence" annotations="5"/>
 <frameAnnotation frame="Justifying" annotations="1"/>
 <frameAnnotation frame="None" annotations="1"/>
 <frameAnnotation frame="Reliance" annotations="3"/>
</ENTRY>
<ENTRY lemma="afsluiten" pos="v">  /* close, settle, end */
 <frameAnnotation frame="Make_agreement_on_action" annotations="1"/>
 <frameAnnotation frame="Sign_agreement" annotations="1"/>
 <frameAnnotation frame="Locale_closure" annotations="2"/>
 <frameAnnotation frame="Change_accessibility" annotations="1"/>
 <frameAnnotation frame="Activity_finish="11"/>
 <frameAnnotation frame="Process_end" annotations="3"/>
</ENTRY>
```

Figure 3: Example of a lexical entry in the Dutch FrameNet lexicon derived from the corpus annotations.

Using the annotations as input for creating a lexicon has an additional advantage. We will explore whether we can group certain annotations and frames and eliminate errors. By addressing the annotations from a lexical point of view, we can critically assess the annotations.

5. Future Plans

The annotation carried out so far follows a traditional **text-to-data** method, where linguists first collect texts and then annotate it with interpretations, e.g. frames. The process is labor-intensive and the IAA is low as explained above. The annotators have to consider a highly diverse set of texts on very different topics. Since they have to annotate every predicate from the PropBank annotation, sentence-by-sentence, they also have to consider all the FrameNet frames and elements continuously.

In future work, we therefore continue with a **data-to-text** approach, described in more detail in Vossen et al. (2018b). This approach starts from a-priori registrations of events in structured data and provides so-called reference texts that report on these specific events. Starting from structured data that defines what the event is, but also who is involved, when and where it took place, the data-to-text approach guarantees a large variety of texts on similar situations and events from various perspectives. Annotators will consider sets of documents that involve more or less the same frames and elements simultaneously in relation to the same or very similar events.

This data-to-text method has several advantages over a classical text-to-data annotation method: 1) we already have predefined a formal representations of events or incidents,

[3] http://tst-centrale.org/nl/tst-materialen/corpora/sonar-klein-corpus-commercieel-detail

often with information on the time, location and participants without having to rely on error-prone automatic processing of text or labor-intensive manual annotation, 2) we obtain a large variety of texts from different sources, genres and languages that make reference to the same events, likely in very different ways, 3) we do not need to interpret everything that is written in the text but can focus on the text parts that relate to the structured data, 4) we can compare many different pairings of structured data and reporting texts for the same type of events and therefore generalize our observations to the level of frame types, 5) annotators can focus on similar events that share frames and frame elements for many texts, 6) annotators can focus more on the variation in framing of similar events.

As explained in (Vossen et al., 2018a; Vossen et al., 2018c), we used this method to annotate 510 documents for event coreference for the SemEval2018-Task5 *Counting Events and Participants in the Long Tail* (Postma et al., 2018). All the documents report on manually registered gun violence incidents and have been annotated given the structured data on the incident a priori.[4] Annotators mark in the text any reference to the incident as a whole and specific subevents. Table 6 lists the most used expressions for the different event types represented by frames. The table shows a wide range of closely related predicates. Note that some of the references to frames can be very indirect, e.g. *surgery* implies *Experience_bodily_harm* and *funeral* implies *Death*. By starting from similar incidents, we not only expect to cover a wider range of predicate and frames but also provide input for possible frame relations that can be added to FrameNet.

Frame	Most common expressions
Death	dead (305) died (285) killed (283)
Use_Firearm	shooting (680) gunshot (247) went off (72)
Hit_Or_miss	shot (801) shooting (83) struck (46) missed (1)
Incident	accident (57) shooting (260) incident (164) tragedy (11) it (88)
Experience_bodily_harm	wound (175) injured (75) injuries (68) surgery (1)

Table 6: Most common expressions used for frames in the Gun Violence corpus

By complementing the current balanced corpus through this vertical extensions by the data-to-text method, we hope to obtain a good mixture of a corpus that on the one hand strives for representing the diversity of language genres and topics and on the other hand for variation in framing similar events across texts. The data-to-text method is different from FrameNet annotation approaches that start from a specific frame and try to find sentences with related lexical units. The event registries do not come with a selection of frames or lexical units and we expect that the annotation of the related texts may involve a substantial variety of related frames. Obviously, only a restricted range of events are covered by the event registries. As such, we consider this method as complementary to other approaches and hope to learn from the differences in variation across these annotations.

6. Conclusion

In this paper, we described the first steps towards an Open Dutch FrameNet lexicon and annotated corpus. The first contribution of this paper is the description of the current status of the annotation process and lexicon. These annotations consisted of adding FrameNet frames and element annotation to a component of the Dutch SoNaR corpus that was already annotated with PropBank predicates and roles. The corpus in question contains a diverse set of written Dutch texts.

A total of 3,898 verbs covering 1,336 predicate types have been annotated with frames and their arguments with frame elements. Due to the high variety of data and lexical types that had to be considered, inter-annotator agreement was lower than in other studies where annotators focused on more selective data. Agreement was 47% for exact match, 51% when counting frames standing in a heritage relation as correct and 54% when accepting frames standing in any relation. Problems were mainly found in the lack of coverage of FrameNet and in mismatches between frames whose distinction is subtle as also observed by Padó (2007). Overall, a lexicon based 4,755 pairings between lexical units and frames could be derived from our data, covering 671 frames.

The second contribution of the paper is that it proposes to use a new method, the data-to-text method (Vossen et al., 2018b) for creating annotated data with a high variation in framing similar events. We plan to apply this method in future work.

7. Acknowledgements

The work presented in this paper was funded by the Netherlands Organisation for Scientific Research (NWO) via the Spinoza grant, awarded to Piek Vossen in the project "Understanding Language by Machines" and VENI grant 275-89-029 awarded to Antske Fokkens.

8. Bibliographical References

Benešová, V., Lopatková, M., and Hrstková, K. (2008). Enhancing czech valency lexicon with semantic information from framenet: The case of communication verbs. *Proceedings of the 1st International Conference on Global Interoperability for Language Resources*, pages 18–25.

Cybulska, A. and Vossen, P. (2010). Event models for historical perspectives: Determining relations between high and low level events in text, based on the classification of time, location and participants. In *LREC*.

Fokkens, A., Ruigrok, N., Gagenstein, S., van Atteveldt, W., and Beukeboom, C. (2018). Microportrait detection for identifying stereotypes in dutch news. In *LREC-2018, Myazaki, Japan.*

Lenzi, V. B., Moretti, G., and Sprugnoli, R. (2012). Cat: the celct annotation tool. In *LREC*, pages 333–338.

Padó, S. (2007). *Cross-lingual annotation projection models for role-semantic information*. Ph.D. thesis, Saarland University.

Petruck, M. R., Fillmore, C. J., Baker, C. F., Ellsworth, M., and Ruppenhofer, J. (2004). Reframing framenet

[4] https://github.com/cltl/GunViolenceCorpus

data. In *Proceedings of The 11th EURALEX International Congress*, pages 405–416.

Postma, M., Ilievski, F., and Vossen, P. (2018). Semeval-2018 task 5: Counting events and participants in the long tail.

Søgaard, A., Plank, B., and Martinez Alonso, H. (2015). Using frame semantics for knowledge extraction from twitter. *Proceedings of the 29th AAAI Conference on Artificial Intelligence*, pages 2447–2452.

Vossen, P., Postma, M., and Ilievski, F. (2018a). From data to text: Capturing long tail events through microworlds and reference texts. In *LREC-2018, Myazaki, Japan*.

Vossen, P., Ilievski, F., Postma, M., and Roxane, S. (2018b). Don't annotate, but validate: a data-to-text method for capturing event data. In *LREC2018, Myazaki*.

Vossen, P., Postma, M., and Ilievski, F. (2018c). Referencenet: a semantic-pragmatic network for capturing reference relations. In *Global Wordnet Conference 2018, Singapore*.

9. Language Resource References

Baker, Collin. (2008). *FrameNet, present and future*.

De Clercq, Orphée and Hoste, Veronique and Monachesi, Paola. (2012). *Evaluating automatic cross-domain Dutch semantic role annotation*.

De Lacalle, Maddalen Lopez and Laparra, Egoitz and Rigau, German. (2014). *Predicate Matrix: extending SemLink through WordNet mappings*.

Kingsbury, P. and Palmer, M. (2002). From treebank to propbank. pages 1989–1993.

Oostdijk, N. and Reynaert, M. and Monachesi, P. and Noord, G. van and Ordelman, R. and Schuurman, I. and Ghinste, V. van. (2008). *From D-Coi to SoNaR: A reference corpus for Dutch*. Paris, France: ELRA.

Postma, Marten and van Miltenburg, Emiel and Segers, Roxane and Schoen, Anneleen and Vossen, Piek. (2016). *Open Dutch WordNet*.