Human Language Technologies – The Baltic Perspective
K. Muischnek and K. Müürisep (Eds.)

doi:10.3233/978-1-61499-912-6-96

Latvian FrameNet: Cross-Lingual Issues

Gunta NEŠPORE-BĒRZKALNE [1], Baiba SAULĪTE and Normunds GRŪZĪTIS
Institute of Mathematics and Computer Science, University of Latvia

Abstract. This paper reports the lessons learned while creating a FrameNet-annotated text corpus of Latvian. This is still an ongoing work, a part of a larger project which aims at the creation of a multilayer text corpus, anchored in cross-lingual state-of-the-art representations: Universal Dependencies (UD), FrameNet and PropBank, as well as Abstract Meaning Representation (AMR). For the FrameNet layer, we use the latest frame inventory of Berkeley FrameNet (BFN v1.7), while the annotation itself is done on top of the underlying UD layer. We strictly follow a corpus-driven approach, meaning that lexical units (LU) in Latvian FrameNet are created only based on the annotated corpus examples. Since we are aiming at a medium-sized still general-purpose corpus, an important aspect that we take into account is the variety and balance of the corpus in terms of genres, domains and LUs. We have finished the first phase of the FrameNet corpus annotation, and we have collected and discuss cross-lingual issues and their possible solutions. The issues are relevant for other languages as well, particularly if the goal is to maintain cross-lingual compatibility via BFN.

Keywords. Latvian, FrameNet, corpus, cross-lingual, NLU

1. Introduction

Latvian FrameNet is being created within a larger research project "Full Stack of Language Resources for Natural Language Understanding and Generation in Latvian" [1], aiming at a balanced multilayer corpus based on cross-lingually oriented syntactic and semantic representations: UD [2], FrameNet [3], PropBank [4], AMR [5], as well as auxiliary layers for named entity and coreference annotation, which are required for the ultimate construction of AMR graphs.

The broad application area we address by creating the FrameNet corpus is natural language understanding (NLU), while more specific applications are, for instance, abstractive text summarization and knowledge base population, which are required by the industrial partner of our project, Latvian news agency LETA, for the automation of various media monitoring processes. NLU systems rely, explicitly or implicitly, on semantic parsing of text. State-of-the-art semantic parsers, in turn, typically rely on supervised machine learning which requires substantial language resources – syntactically and semantically annotated text corpora. Experience that our research group has gained through developing semantic parsers [6,7] and semantics-based text generators [8,9], by combin-

[1]Corresponding Author: Gunta Nešpore-Bērzkalne, Artificial Intelligence Laboratory, Institute of Mathematics and Computer Science, University of Latvia, Raiņa bulv. 29, Riga, LV-1459, Latvia; E-mail: gunta.nespore@ailab.lv.

ing machine learning and grammar engineering approaches, has convinced us that both FrameNet and AMR have a great potential to establish as powerful and complementary semantic interlinguas which can be furthermore strengthened and complemented by other multilingual representations.

We aim to acquire a balanced and representative medium-sized corpus of Latvian: around 10,000 sentences annotated at all the above mentioned layers, including FrameNet. To ensure that the corpus is balanced not only in terms of text genres and writing styles but also in terms of LUs, a fundamental design decision is that the text unit is an isolated paragraph. Paragraphs are manually selected from a balanced 10-million-word text corpus: 60% news, 20% fiction, 7% academic texts, 6% legal texts, 5% spoken language, 2% miscellaneous.

As for the LUs, 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 LUs, i.e., word senses w.r.t. FrameNet frames, will be considerably larger. We assume that the corpus will prove to be balanced also w.r.t. nominal LUs.

2. Latvian FrameNet

There are many FrameNet-like language resources, some of them evolved independently of the original Berkeley FrameNet (BFN) [10]. Approaches used in building framenets differ, and it is often the case that rather different inventories of abstract semantic frames are defined and used, based on the language data and language specifics. Most projects, however, try to reuse the original BFN frames as far as possible, before introducing additional or language-specific frames. Yet another approach is to translate an existing FrameNet-annotated corpus from a source language to a target language. Nevertheless, the different projects try to keep an eye on how their frame inventories compare with frames created in BFN.

Apart from general-purpose framenets, some projects focus on restricted domains, e.g. French FrameNet focuses on four notional domains: verbal communication, commercial transactions, cognitive stance, and causality [10], FrameNet Brazil – on tourism and sports [11], while a previously developed domain-specific Latvian FrameNet focuses on the media monitoring use case, addressing only 25 modified BFN frames [12].

The annotation of the general-purpose Latvian FrameNet is based on the latest BFN frame inventory (v1.7). We stick to the BFN frames in order to reuse the BFN frame hierarchy and other inter-frame relations, as well as semantic types of frame elements (FE), and the definitions of frames and FEs in general. Another reason for BFN compatibility is to facilitate use cases that require cross-lingual semantic parsing. Note that Latvian FrameNet itself, as well as its inventory of LUs, is being acquired implicitly by annotating corpus examples – frame instances.

3. Annotation process

The annotation of Latvian FrameNet corpus examples is done on top of UD trees [13] using the annotation tool WebAnno [14] which supports a centralized web-based anno-

tation workflow. As for the FEs, the underlying UD tree allows for selecting only head nodes while annotating FEs (see Figure 1). The full span of an FE can be acquired automatically by traversing the respective UD subtree.

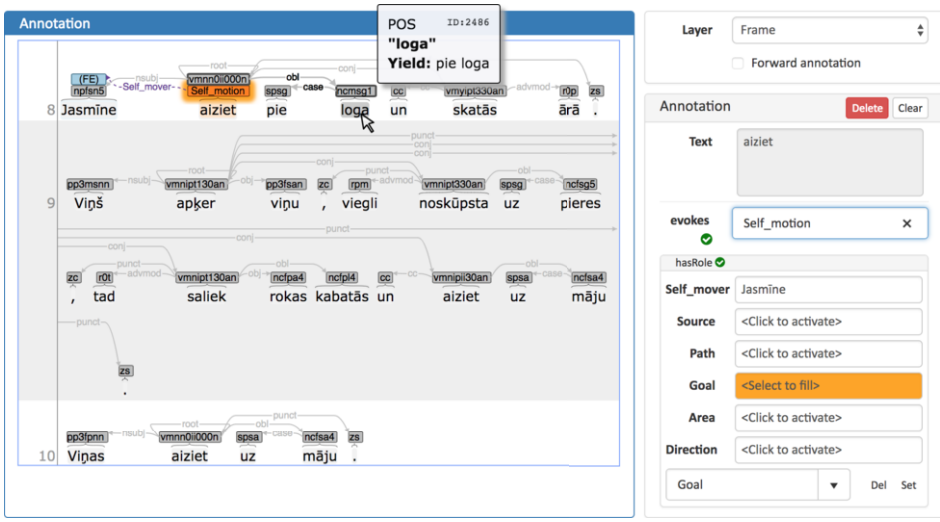

Figure 1. A screenshot of the annotation mode in WebAnno.

First, an annotator selects a target word that evokes a frame and specifies the frame being evoked. We have configured WebAnno, so that the most likely frame (depending on the lemma of the target word) appear at the top of the drop-down list of BFN frames. Then, the annotator specifies the head nodes of the core FEs which are specific to the frame and are selected from a predefined template adjusted to the selected frame.

When one annotator has finished her set of corpus examples, another annotator (i.e., curator) approves the annotations (see Figure 2). In the curation mode, the underlying UD trees are hidden, so that the curator can focus solely on FrameNet annotations, consulting the UD tree only if necessary. Note that in case of inter-annotated corpus examples, disagreement between annotators is highlighted by WebAnno (see Sentences 9 and 10 in Figure 2).

Unlike in the recent shared annotation task of the Multilingual FrameNet initiative [15], we do not intentionally conduct full-text annotation. Instead, we follow a concordance approach: we annotate frame instances target word by target word instead of document by document (and then sentence by sentence). Such approach increases the annotation consistency. Many sentences, however, will become fully annotated after merging annotation sets of the same sentence from different concordances. The final result of merged annotation sets is illustrated in Figure 3.

Another difference from BFN and most other framenets is that we systematically annotate only core FEs (which characterize and define the frame). Additionally, we systematically annotate only two non-core FEs: TIME and PLACE which are important in many information extraction use cases.

Otherwise we strictly follow a corpus-driven approach: LUs in Latvian FrameNet are created only based on the annotated corpus examples. Moreover, Latvian FrameNet

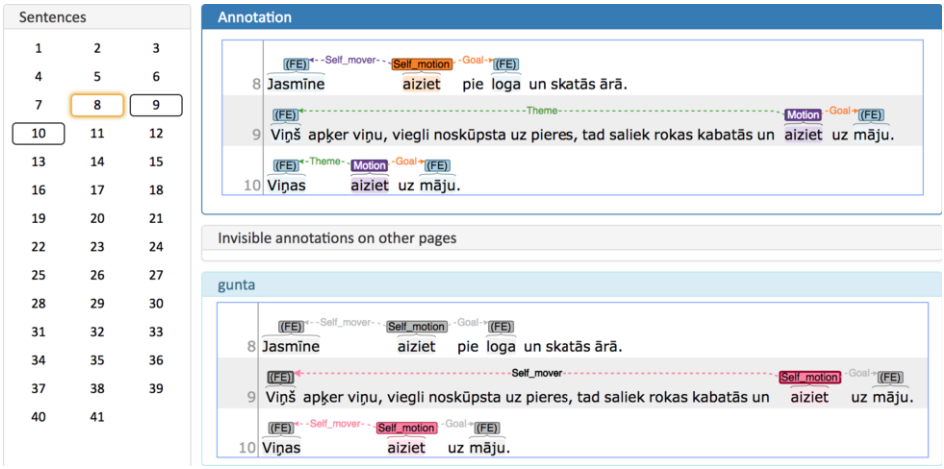

Figure 2. A screenshot of the curation mode in WebAnno.

annotation is done on the top of the underlying UD layer. The annotation of frames and FEs is thus guided by the dependency structure of a sentence, instead of the phrase structure. More details on the annotation process are described in [13].

Besides verbs as LUs, we have also started the annotation of frequent nominalizations as illustrated by the frame DENY_OR_GRANT_PERMISSION in Figure 3. This instance of the frame is evoked by the noun *atļauja* 'permission'.

Current statistics of the Latvian FrameNet corpus:

- 778 different target words (lexemes);
- 7,024 annotation sets (cf. 174,022 annotation sets in BFN);
- 432 different frames (cf. 1,087 lexical frames in BFN);
- 1,421 different LUs (cf. 8,393 LUs with annotated examples in BFN).

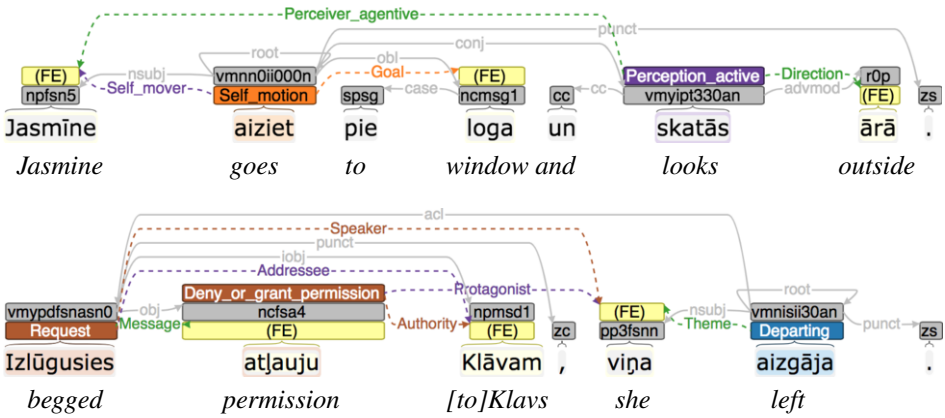

Figure 3. Full-text FrameNet annotation, acquired by merging annotation sets of all separately annotated frame instances. The FrameNet layer is complemented by the UD layer.

4. Cross-lingual issues

At the time of writing, the first stage of the Latvian FrameNet annotation is completed: verbs with 3 or more usage examples in the current version of Latvian UD Treebank (v2.2; 7,703 sentences) are mapped to BFN frames (thus, creating LUs), and their occurrences are annotated (thus, creating annotation sets).

Like other FrameNet projects that reuse BFN frames [10], we are facing certain difficulties to provide mapping between BFN frames and Latvian verbs, in order to create an LU. These difficulties that arise from lexical differences between languages were expected and have to be addressed. We have identified and classified the issues into several groups along with one or more solutions.

4.1. Frames not covered by BFN

Some word senses or concepts are simply not covered by the current version of BFN, for example: *veltīt* 'to devote', *atjaunot* 'to update', *svinēt* 'to celebrate', *balsot* 'to vote', *rēķināt* 'to calculate', *atgūt* 'to recover, to get back', *iepazīstināt* 'to introduce', *ziedēt* 'to blossom', *sapņot* 'to dream', *rakt* 'to dig'. These cases are not specific to the Latvian language or culture.

Alternative solutions.

1. Define and add new frames to the BFN frame hierarchy. A global solution would be a methodology and a procedure for proposing and considering new frames in the common BFN frame inventory, preferably in the scope of the Multilingual FrameNet initiative [10] or in another coordinated way to obtain a result compatible and reused among different languages. Before a global solution is established, we have to use a local solution: to define a necessary new frame by inheriting it from the closest general BFN frame, or to leave (temporarily) the occurrences of a target word without providing annotation sets.
2. Use a more general frame when an appropriate frame cannot be found.[2] This often means the use of a very general and, thus, not very informative frame, like the INTENTIONALLY_ACT frame for the verbs *rakt* 'to dig' and *iepazīstināt* 'to introduce'. For some verbs, it is difficult to find even a general frame; for example, for the verb *ziedēt* 'to blossom'.

Chosen solution. After identifying all or most of the missing frames, we will introduce a minimal set of BFN-inherited frames required to complete the annotation of Latvian FrameNet. A similar approach of inheriting new frames from BFN has been applied also for Swedish FrameNet [16].

4.2. Substantial differences in word sense splitting between Latvian and English

In some cases, there is no lexical correspondence between concepts in English and Latvian: the meaning of a Latvian verb corresponds to the meaning of an English phrase, or vice versa.

[2] A temporal solution suggested by the coordinators of the Multilingual FrameNet initiative.

4.2.1. The sense of a Latvian verb is more specific

A Latvian verb stands for a concept that is expressed by a phrase in English, i.e., the sense of the Latvian verb is more specific compared to the respective sense of the closest English verb or compared to the definition of the closest BFN frame. For example: *pārdomāt* 'to change one's mind' – BFN frames related to thinking (OPINION, COGITATION) do not fit this verb sense, since they do not involve the concept of change that is present in the meaning of the Latvian verb. The more general BFN frame CAUSE_CHANGE does not fit for similar reasons. Equally, we have not found a good mapping for verbs like: *atvadīties* 'to say goodbye', *maldīties* 'to be wrong', *saņemties* 'to pull oneself together', *rūpēties* 'to take care of', *pārņemt* 'to take over'.

Alternative solutions. Same as in 4.1.

Chosen solution. Same as in 4.1.

4.2.2. The sense of a Latvian verb is more general

An English verb stands for a concept that is expressed by a phrase in Latvian. For example: *lasīt lekciju* 'to lecture' ('to read a lecture'), *aiziet pensijā* 'to retire' ('to go on pension'), *krist ģībonī* 'to faint' ('to fall into unconsciousness'), *likt lietā* 'to use' ('to put in use'), *spert soli* 'to step' ('to take/kick a step'), *iet prom* 'to leave' ('to go away'), *uzlikt par pienākumu* 'to oblige' ('to put as a duty'), *taisīt vaļā* 'to open' ('to make open'). Some of them (e.g. *likt lietā* 'to use') are multi-word expressions and are not problematic, since FrameNet supports multi-word LUs, but others (e.g. *lasīt lekciju* 'to lecture') are regular phrases that normally are not treated as multi-word units in Latvian.

Alternative solutions.

1. Treat such verb phrases as multi-word LUs, even if it is arguable from the lexicographic perspective.
2. Use the closest BFN frame if possible, e.g. if the direct object or other complement of the target verb can be annotated as an FE (preferably, as a core FE). It would work for verb phrases like *aiziet pensijā* 'to retire', if going on pension could be seen as evoking the MOTION frame where pension is GOAL,[3] but not for verb phrases like *spert soli* 'to step' ('to take/kick a step'), as the SELF_MOTION frame has no FE that would fit the direct object *soli* 'a step'.

Chosen solution. The first option, if the second one leads to evoking a frame that contradicts annotator's intuition and/or significantly differs from the frame that covers the corresponding concept in English.

4.2.3. Different semantic elements between Latvian and English verb senses

The semantic elements are different between Latvian and English verb senses. For example, *braukt* 'to move using a vehicle': the meaning 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]'). From the Latvian perspective, this verb evokes the frame USE_VEHICLE (a non-lexical frame in BFN), instead of frames RIDE_VEHICLE or OPERATE_VEHICLE.

Solution. Use a non-lexical BFN frame. A frame that according to the BFN corpus is non-lexical might have associated LUs in another languages [17].

[3] A more appropriate frame, however, would be QUITTING.

4.3. A concept in Latvian and English is expressed by different parts of speech

A meaning of the Latvian verb roughly corresponds to a nominal BFN frame (mostly evoked by adjectives and nouns as target words). For example: *klusēt* 'to be silent' (VOLUBILITY), *piedzerties* 'to get drunk' (INTOXICATION), *padoties* 'to be good at' (EXPERTISE), *salt* 'to be cold' (SUBJECTIVE_TEMPERATURE), *nogurt* 'to get tired' (BIOLOGICAL_URGE).

Solution. Use nominal BFN frames for representing Latvian verb senses. Since BFN frames are often evoked by LUs of different parts of speech (differences are manifested in the description of the syntactic realizations of FEs), adding a verbal LU to a frame with no verbal LUs so far is an acceptable solution.

5. Conclusion

By strictly sticking to the BFN frame inventory, we have finished the first stage of Latvian FrameNet annotation. The annotation progress has been relatively rapid due to the UD-based approach which allows for annotating only the syntactic roots of FEs instead of whole text spans, and due to the verb by verb annotation methodology. Meanwhile, we have identified several groups of problematic mappings between Latvian verb senses and BFN frames.

We have ascertained that the inventory of BFN frames is not sufficient even to cover all senses of the most frequently used Latvian verbs. Therefore we have proposed solutions for tackling such cases. These solutions have already been partly implemented in Latvian FrameNet; the rest – requiring to introduce new inherited frames – will be implemented in the upcoming annotation stage.

The annotated Latvian FrameNet data is available on GitHub.[4]

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 (*Full Stack of Language Resources for Natural Language Understanding and Generation in Latvian*) and No. 1.1.1.2/VIAA/1/16/188 (*From Abstract Meaning Representation to Natural Language Sentence and Coherent Text Generation*).

References

[1] N. Gruzitis, L. Pretkalniņa, B. Saulīte, L. Rituma, G. Nespore-Berzkalne, A. Znotiņš and P. Paikens, 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, 2018, pp. 4506–4513.

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

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

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

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

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

[6] G. Barzdins, D. Gosko, L. Rituma and P. Paikens, 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)*, Reykjavik, Iceland, 2014, pp. 4476–4482.

[7] G. Barzdins and D. Gosko, RIGA at SemEval-2016 Task 8: Impact of Smatch extensions and character-level neural translation on AMR parsing accuracy, in: *Proceedings of the 10th International Workshop on Semantic Evaluation*, San Diego, California, 2016, pp. 1143–1147.

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

[9] N. Gruzitis, D. Gosko and G. Barzdins, 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*, Vancouver, Canada, 2017, pp. 924–928.

[10] L. Gilardi and C. Baker, Learning to Align across Languages: Toward Multilingual FrameNet, in: *International FrameNet Workshop 2018: Multilingual FrameNets and Constructicons*, Miyazaki, Japan, 2018, pp. 13–22.

[11] A. Costa, M. Gamonal, V. Paiva, N. Marção, S. Peron-Corrêa, V. Almeida, E. Matos and T. Torrent, FrameNet-Based Modeling of the Domains of Tourism and Sports for the Development of a Personal Travel Assistant Application, in: *Proceedings of the International FrameNet Workshop 2018: Multilingual FrameNets and Constructicons (IFNW)*, Miyazaki, Japan, 2018, pp. 6–12.

[12] G. Barzdins, FrameNet CNL: A knowledge representation and information extraction language, in: *Controlled Natural Language*, Lecture Notes in Computer Science, Vol. 8625, Springer, 2014, pp. 90–101.

[13] N. Gruzitis, G. Nespore-Berzkalne and B. Saulite, Creation of Latvian FrameNet based on Universal Dependencies, in: *Proceedings of the International FrameNet Workshop 2018: Multilingual FrameNets and Constructicons (IFNW)*, Miyazaki, Japan, 2018, pp. 23–27.

[14] R. Eckart de Castilho, E. Mújdricza-Maydt, S.M. Yimam, S. Hartmann, I. Gurevych, A. Frank and C. Biemann, 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*, Osaka, Japan, 2016, pp. 76–84.

[15] T. Torrent, M. Ellsworth, C. Baker and E. Matos, The Multilingual FrameNet shared annotation task: A preliminary report, in: *Proceedings of the International FrameNet Workshop 2018: Multilingual FrameNets and Constructicons (IFNW)*, Miyazaki, Japan, 2018, pp. 62–68.

[16] K. Friberg Heppin and M. Toporowska Gronostaj, The Rocky Road towards a Swedish FrameNet - Creating SweFN, in: *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC)*, Istanbul, Turkey, 2012, pp. 256–261.

[17] J. Ruppenhofer, M. Ellsworth, M.R.L. Petruck, C.R. Johnson, C.F. Baker and J. Scheffczyk, *FrameNet II: Extended Theory and Practice*, International Computer Science Institute, Berkeley, California, 2016.