Towards Consistent UMR Annotation of Deverbal Nouns: Evidence from Czech and Latin

Hana Hledíková, Federica Gamba, Marketa Lopatkova, Jan Štěpánek
Charles University, Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics
Malostranské nám. 2/25, 118 00 Prague 1, Czechia
{hana.hledikova, gamba, lopatkova, stepanek}@ufal.mff.cuni.cz

Abstract

Deverbal nouns pose challenges for semantic annotation frameworks that aim to represent event structures consistently across lexical categories. This paper examines problematic phenomena in the annotation of deverbal nouns in Czech and Latin within the Universal Meaning Representation (UMR) framework, addressing both manual graph construction and rule-based automatic conversion from existing resources. Current UMR guidelines lack operational criteria for deciding when a noun should be treated as an eventive concept, particularly in the absence of a PropBank-like lexicon with sufficient nominal coverage. We therefore propose practical annotation principles: deverbal nouns denoting events (such as *učení* 'teaching'), results of events (*řešení* 'solution'), or event participants (*učitel* 'teacher') should be related to underlying event concepts (represented as verbs in their particular senses, i.e., *učit-001* 'to teach', *vyřešit-001* 'to solve', and *učit-001* 'to teach', respectively), while other deverbal nouns should remain unrelated to respective events (such as *učebna* 'teaching room'). To reduce inter-annotator variation, we further suggest systematic strategies for selecting verbal labels, including the use of light-verb constructions, synonymous verbs, and a preference for imperfective verbs in Czech aspectual pairs. For automatic conversion, we outline a rule-based approach that combines multiple lexical resources and frequency-based heuristics to identify corresponding verb senses. Our findings provide guidelines for more consistent UMR annotation across languages.

Keywords: UMR, event nouns, Czech, Latin, Prague Dependency Treebank, Latin Dependency Treebank, automatic conversion

1. Introduction

Uniform Meaning representation (UMR, see esp. van Gysel et al., 2021; Bonn et al., 2024; Bonn et al., 2026)[1] is a framework designed to capture the semantic content of a text in any language. UMR is based on Abstract Meaning Representation (AMR, Banarescu et al., 2013; Wein and Bonn, 2023), originally developed primarily for English with its rich linguistic resources, extending this approach to make it applicable to other languages, particularly those with rich morphology and limited linguistic resources.

In this vein, UMR has been employed to represent Czech and Latin, languages that have relied so far on a sophisticated dependency-based description of deep syntax (Sgall et al., 1986). In addition to preparing a sample of manually annotated data for both languages, two existing corpora—PDT-C (Hajič et al., 2020; Hajič et al., 2024) for Czech and LDT[2] for Latin (Bamman and Crane, 2006; Passarotti, 2014; Gonzalez Saavedra and Passarotti, 2014)—have been leveraged to create a large automatically converted UMR dataset for Czech and Latin (Štěpánek et al., 2025a; Štěpánek et al., 2025b).

During the work on the dataset, *deverbal nouns* and *deverbal adjectives* were identified as problematic phenomena for UMR for two main reasons: first, it is often unclear what the "correct" UMR annotation should be, as multiple analyses may be plausible (Lopatková et al., 2025b); second, even when a preferred annotation can be determined, producing it reliably through automatic conversion remains challenging.

1.1. Need of a PropBank-like lexicon for events

A fundamental requirement of UMR annotation is the distinction between entities and events, the latter of which can be further divided into states and processes. Events are prototypically expressed through predication, i.e., by verbs; however, they may also be realized by event-denoting nominals or adjectives.

In this respect, English UMRs are based on the PropBank lexicon (Palmer et al., 2005; Pradhan et al., 2022),[3] which provides 'frame files' with detailed information on event participants and argument structure; the frames are populated not only with verbs but also with participles, light-verb constructions, event-denoting nouns, and event-denoting adjectives; cf. the 'break.01' frame (∼

[1] https://umr4nlp.github.io/web/index.html

[2] https://itreebank.marginalia.it/

[3] https://github.com/propbank/

Proceedings of The Seventh International Workshop on Designing Meaning Representations (DMR 2026) @ LREC 2026, pages 65–80
11 May 2026.

break, cause to not be whole), which lists break (v.), break (n.), breaking (n.), make_break (l.), broken (j.).

For Czech, two datasets can be used: (i) the PDT-Vallex lexicon (Hajič et al., 2003; Urešová et al., 2021), developed as a resource for valency annotation in PDT-C, and (ii) the Czech part of the SynSemClass ontology (Urešová et al., 2025a; Urešová et al., 2025b). While entries for verbal predicates were successfully (semi)automatically converted to PropBank-like frames (Hajič et al., 2024), the coverage of event-denoting nouns and adjectives is very limited in these resources; therefore, they have not been processed so far.

For Latin, two valency lexicons have been developed over the years: (i) Latin Vallex 1.0 (Passarotti et al., 2016), consisting of approximately 2,500 valency frames, grounded in the tectogrammatical layer of LDT and ITTB (Passarotti, 2019),[4] but not differentiating frames on semantic grounds; and (ii) Latin Vallex 2.0 (Mambrini et al., 2021), which expands coverage to over 45,000 valency frames and links entries to WordNet synsets via the LiLa Knowledge Base (Passarotti et al., 2020), but does not provide cross-references to LDT. Additionally, its usability is limited due to the absence of illustrative examples and the frequent inclusion of highly similar, or even identical, frames. Since these two resources are essentially independent of each other, with little to no overlap beyond some common lemmas, the two valency lexicons have been (partially) combined into Vallex4UMR,[5] suitable for UMR annotation but covering only a subset of LDT. These resources are not restricted to verbs, but also encompass nouns and adjectives.

In this paper, we focus on the challenges that deverbal nouns pose for UMR annotation in Czech and Latin, two languages with rich derivational morphology, but without any PropBank-like valency lexicon that would systematically include nouns along with verbs and could therefore be readily exploited for automatic conversion. Section 2 introduces the issues connected with determining the appropriate UMR representation of deverbal nouns in Czech and Latin, Section 3 focuses on the issues for automatic conversion of existing data formats into UMR and presents some preliminary findings on how additional data resources can be utilized to address them. Section 4 closes the paper with a summary of the findings and brief concluding remarks.

[4]Consequently, each predicate occurring in the tectogrammatical layer is associated with a corresponding valency frame recorded in Vallex 1.0.

[5]https://github.com/fjambe/Vallex4UMR

2. Deverbal nouns with unclear UMR representation

2.1. Events vs. non-events denoted by deverbal nouns

As we have described in Section 1.1, eventive concepts in UMR are not necessarily realized by verbs. Nouns that are derived from verbs and also denote events (i.e., event nouns) are annotated using an eventive concept labeled with the corresponding verb-sense in the reference valency lexicon (cf. the noun *učení* 'teaching' in example 1). Furthermore, derived nouns that denote a participant in an event can also be annotated using an eventive concept in combination with an inverse participant relation, because the eventive meaning of the verb is also clearly referred to by the deverbal noun; cf. agent nouns such as *učitel* 'teacher' in example 2.

Deverbal nouns can also denote other kinds of participants and circumstances; cf. the noun *učivo* 'teaching material; curriculum', which denotes the second argument of *učit-001* 'to teach', or the noun *učebna* 'teaching room', which denotes the location. In such examples, it is less clear whether the UMR representation should still use the eventive concept (although the noun clearly refers to it); rather, a simple entity concept labeled with the noun's lemma seems more appropriate.

In general, the guidelines[6] do not offer testable criteria to identify nouns that should refer to events. There are nouns that are not derived from verbs, but are prototypical participants of events; cf. the primary noun *žák* 'pupil', which is also conceptually associated with an event of teaching (cf. the potentially possible representation in 3), but we would use an entity concept to represent it (cf. 4). Even nouns derived from verbs can be prototypically associated with multiple different events; cf. the noun *jídlo* 'food', which is derived from *jíst* 'to eat', but there is no principled reason not to annotate it with reference to a different event in which it also participates, such as *vařit* 'to cook'. Although the fact that a noun is derived from a verb is a certain indication that may lead to preferring a certain eventive concept, morphology is language-specific and should not, in principle, be the criterion for the UMR representation. In the absence of a reference valency lexicon that would list the nouns that are linked to a particular verbal frame, some testable criteria need to be specified to choose the appropriate representation–either a simple entity concept or a corresponding eventive concept.

(1) *učení* 'teaching'

```
(s1u1 / učit-001 'to teach'
    ...)
```

[6]https://github.com/umr4nlp/umr-guidelines/blob/master/guidelines.md

(2) *učitel* ‘teacher’

```
(s1p1 / person
        :refer-number singular
        :ARG0-of (s1u1 / učit-001
                 'to teach'
                 ...))
```

(3) *žák* ‘pupil’

```
(s1p1 / person
        :refer-number singular
        :ARG1-of (s1u1 / učit-001
                 'to teach'
                 ...))
```

(4) *žák* ‘pupil’

```
(s1z1 / žák 'pupil'
        :refer-number singular)
```

Essentially, we can view deverbal nouns that denote events as differing from verbs simply in the “information packaging” (Croft, 2001, 2022): the same semantic type (a process) is presented using the referential function rather than the predication function, which is reflected in the use of a different part of speech—a noun instead of a verb. In Kuryłowicz (1936), this kind of noun formation is discussed under the term “syntactic derivation”, in contrast to “lexical derivation” which also involves a change in the kind of concept that is denoted by the derived word vs. the base word and is exemplified by the nouns *učitel* ‘teacher’, *učivo* ‘teaching material; curriculum’ or *učebna* ‘teaching room’. In nouns of this second type, we suggest formulating clear criteria for manual annotation to decide when to use an eventive concept, such as limiting the eventive label to cases where the noun denotes an argument of the verb, and not another kind of circumstance that would correspond to a non-argument semantic relation (such as `:place-of` in the case of *učebna* ‘teaching room’). The first type of nouns (i.e., those created by “syntactic derivation”) should clearly be represented as an event, but there is still the issue of identifying the appropriate verb-sense that should be used in the eventive concept label in the absence of a PropBank-like lexicon.

2.2. Choosing the appropriate verb entry

Even if choosing the verb does not seem problematic at first look, there are certain properties of derivation which make the decision not obvious in some nouns. Derivation is traditionally characterized as only partially predictable and partially productive (cf. e.g. Tuggy, 1985). In contrast to inflection, where non-existent forms for a given paradigm cell are rather unexpected, unavailable forms for a given derivational meaning are frequently found in derivation (Stump, 2019). While most nouns with an eventive meaning do have a corresponding verb (see, e.g., *boj* and *pugna* ‘fight’ in Table 1), it is also possible to find nouns for which a corresponding verb does not exist. The reason is often lexical borrowing (cf. *akvizice* ‘acquisition’ in Table 1, where only the noun was borrowed into Czech but the verb was not), but it also concerns native vocabulary (cf. *krok* ‘step’ and *iter* ‘journey’ in Table 1). This poses a problem for the UMR annotation: if the noun is not included in the reference valency lexicon, a corresponding verb-sense has to be identified, and it is not clear what to do for nouns without a corresponding verb available in the language.

Light-verb constructions. In such cases, a light-verb construction can be used instead of the missing corresponding verb, e.g., *provést akvizici* ‘to carry out an acquisition’, *udělat krok* ‘to make a step’. Because the Czech reference valency lexicon PDT-Vallex (Urešová et al., 2021) (cf. Section 1.1) also contains light-verb constructions, it is possible to refer to those entries in the annotation. For the noun *akvizice*, the concept `provést-akvizici-004` can be chosen, corresponding to the PDT-Vallex frame *provést* `v41hrmF` ‘to carry out an action’. Problems are posed by event nouns that do not have any light-verb construction available in PDT-Vallex (e.g., *publicita* ‘publicity’, where a light-verb construction such as *dělat publicitu* ‘to do publicity’ is not available in the lexicon) and, conversely, by event nouns that have multiple possible light-verb constructions available in PDT-Vallex (e.g., *učinit krok* ‘to carry out a step’ and *udělat krok* ‘to make a step’, which are both synonymous light-verb constructions for the noun *krok* ‘step’ and which are both included in PDT-Vallex). Both options are essentially correct and both can be used, the issue is only with introducing unnecessary inter-annotator disagreements into the data in case it is evaluated.

For Latin, none of the available valency lexicons (cf. Section 1.1) contains entries for light-verb constructions. This presents a challenge for event nouns that lack a corresponding single verb (e.g., *iter* ‘journey’, which is conventionally used in the light-verb construction *iter facere* ‘to make a journey’). Unlike in Czech, where light-verb constructions are included in PDT-Vallex and can be referenced in annotation, Latin event nouns of this type cannot be linked to an existing light-verb frame, which complicates the annotation process.

Aspectual pairs. Multiple synonymous light-verb constructions are not the only type of situation where there are several possibilities to choose from when annotating an event noun. A frequently occurring example in Czech is associated with the characteristics of its morphological system, namely the verbal category of grammatical aspect. Many Czech verbs form so-called aspectual pairs, i.e., pairs of corresponding verbs with the same root and

Czech	
action noun	**verb(s)**
boj 'fight'	*bojovat* 'to fight'
adopce 'adoption'	*adoptovat* 'to adopt'
akvizice 'acquisition'	—
krok 'step'	—
prodávání 'selling'	*prodávat* 'sell (imperf.)'
prodej 'sale'	*prodat* 'sell (perf.)', *prodávat* 'sell (imperf.)'
řešení 'solution'	*řešit* 'solve (imperf.)', *vyřešit* 'solve (perf.)'

Latin	
action noun	**verb(s)**
pugna 'fight'	*pugno* 'to fight'
acquisitio 'acquisition'	*acquiro* 'to acquire'
adventus 'arrival'	*advenio* 'to arrive'
iter 'journey'	—
usus 'use'	*utor* 'to use'
venditio 'selling/sale'	*vendo* 'to sell'

Table 1: Pairs of an event noun and its corresponding verb(s) where available.

basic lexical meaning, but differing in grammatical aspect—one is imperfective (imperf.) and one is perfective (perf.); cf. *prodat* 'sell (perf.)' – *prodávat* 'sell (imperf.)' and *řešit* 'solve (imperf.)' – *vyřešit* 'solve (perf.)' in Table 1. Each of the two verbs is treated as a separate lexeme in both traditional dictionaries and in PDT-Vallex.

Some deverbal nouns, specifically those formed by the suffix *-ní/tí*, can explicitly preserve information about the grammatical aspect of the base verb in their form, and it is therefore clear which verb to use for annotating the eventive concept (cf. *prodávání* 'selling' in Table 1). However, nouns formed via other processes, such as conversion (but also some *-ní/tí* nouns, cf. *řešení* 'solution' in Table 1), usually do not preserve aspectual information and both verbs from the aspectual pair could be chosen (cf. *prodej* 'sale' in Table 1). Sometimes the sentential context disambiguates the perfective vs. imperfective reading (cf. examples 5 and 6), but it is often the case that both verbs from the aspectual pair could be chosen in a particular occurrence—grammatical aspect is simply unspecified. In spite of this, only one of the verbs in the aspectual pair has to be chosen, because they are treated as separate lexemes in the reference valency lexicon.

(5) *Prohlásil, že vypracoval* ***řešení****.*
'He announced that he worked out a **solution**.'
```
(s1t1 / thing
    :refer-number singular
    :result-of (s1v1 / vyřešit-001
                'to solve (pf.)'
                ...))
```

(6) *při* ***řešení*** *matematické úlohy* (...)
'when **solving** a mathematical problem' (...)
```
(s1r1 / řešit-001
        'to solve (impf.)'
        ...)
```

Polysemy. The polysemy of derivational processes may also be problematic for manual annotation. Because a single derived noun may have different types of semantic relations to its base verb depending on the particular sense, this means that the annotator has to determine which sense was used in each particular context. This is difficult especially in nouns with abstract meanings; for instance, the aforementioned noun *řešení* 'solution' was shown to have both the eventive meaning (the action of solving as in 6) and the resultative meaning (the result of solving as in 5). However, in some contexts, these two meanings are difficult to tell apart, sometimes to such degree that the word may be considered ambiguous (cf., e.g., the sentence *Ale situace nedovolovala jiné řešení* 'The situation did not allow for a different solution'). A very similar issue arises in Latin with deverbal nouns such as *solutio* 'solution' or *venditio* 'selling/sale'. For instance, *solutio* can denote either the act of solving or the resulting solution. The noun *venditio* likewise exhibits this polysemy, displaying an eventive meaning, as in *Venditio est rei suae in alium translatio* 'A sale (act of selling) is the transfer of one's property to another', and a resultative meaning, as in *Antequam venditio transferatur* 'Before the sale (object sold) is transferred'.

It is especially the eventive and resultative meanings that are typically difficult to distinguish. We suggest annotating resultative nouns with an eventive concept in combination with the `:result-of` relation, as it is very close to the eventive meaning, but we expect that the choice between the eventive and resultative reading is a potential source of inter-annotator variation.

A decision tree that supports annotators' decisions is provided in Appendix A.

3. Deverbal nouns in the automatic conversion

When it comes to the automatic generation of UMR graphs from available corpora—PDT-C for Czech and LDT for Latin (see Section 3.1 for the basic characteristics of the corpora; the conversion procedure is discussed by Štěpánek et al., 2025a; Lopatková et al., 2025b)—the main issue is identifying whether a particular noun occurrence should be represented using an eventive concept or not.

The eventive representation is appropriate for those noun occurrences that have an eventive meaning, for those with a resultative meaning, and for those that denote an argument in an event. As already mentioned, derivational morphology can be a guide in this regard, but as we have seen, most derivational processes are polysemous.

Although neither PDT-C nor LDT provides explicit semantic annotation for nouns, both resources annotate certain nouns' syntactic dependents with participant roles (e.g., Actor, Patient). However, this is not done in a systematic way in PDT-C. In LDT, some nouns are also associated with a Vallex 1.0 valency frame. This can be interpreted as an indication that the noun has an eventive meaning in that particular occurrence.

When the eventive representation is chosen, in the absence of an entry for the noun-sense in the reference valency lexicon, a base verb for the given derived noun must be identified. It is not particularly difficult since derivational resources exist both for Czech and Latin (cf. Section 3.1), making it possible to automatically identify the verb from which a noun was derived in the majority cases. However, it is necessary to identify not only the verb, but also the particular sense of the verb from which the noun was derived (in case the verb is polysemous).

Another step that has to be carried out once the correct eventive concept (i.e., a particular sense of the verb) is identified is to map the original noun's dependents onto the argument roles of the eventive concept. When the noun's dependents are annotated using participant roles, the mapping can be carried out using the same procedure that is applied to verbs and their arguments (Hajič et al., 2024). Example 7 shows the changes that are necessary to convert the noun *debata* 'debate' modified by the possessive pronoun *jejich* 'their' and the prepositional phrase *o systémech* 'about systems' from the tectogrammatical layer in PDT-C to the UMR representation. Similarly, example 8 illustrates for Latin how the noun *studium* 'zeal', together with the possessive modifier *suum* 'his' and the prepositional complement *in rem publicam* 'toward the state', is transformed from its tectogrammatical representation in LDT into the corresponding UMR structure.

In cases where the dependents are not annotated using participant roles in the original data format, the task of identifying the participant roles is more difficult, but morphological features such as preposition, nominal case, or pronoun type can be used to estimate the participant role in certain cases. In the Czech example, the preposition *o* 'about' + the locative case in *o systémech* 'about systems' indicates that this is the second argument (`ARG1`) of the verb *debatovat* 'to debate'. However, many morphological forms are ambiguous and cannot be used to decide the type of argument with certainty.

(7) *jejich debata o systémech*
'their debate about systems'

a. PDT-C (tectogrammatical layer)

```
(debata 'debate'
    :ACT oni 'they'
    :PAT systém 'system'
    ... )
```

b. UMR

```
(s1d1 / debatovat-001
    'to debate'
    :ARG0 (s1p1 / person
        :refer-number plural
        :refer-person 3)
    :ARG1 (s1s1 / systém 'system'
        :refer-number plural)
    ... )
```

(8) *studium suom in rem publicam*
'his zeal for the state'

a. LDT (tectogrammatical layer)

```
(studium 'zeal'
    :ACT is 'he'
    :PAT (res 'thing'
        :RSTR publicus 'public')
    ... )
```

b. UMR

```
(s2s1 / studeo-001
        'to devote oneself to'
    :ARG0 (s2p1 / person
        :refer-number singular
        :refer-person 3)
    :ARG1 (s2r1 / res 'thing'
        :mod (s2p2 / publicus
                     'public')
        :refer-number singular)
    ... )
```

For nouns that denote an argument of the verb or have the resultative meaning, the graph structure must also be substantially changed. Compare, for instance, the annotation of the phrase in 9: while in PDT, the agent noun *učitel* 'teacher' is the parent node of two dependent nodes *můj* 'my' and *dějepis* 'history', the UMR structure is more complex: an abstract concept `person` serves as the parent of a new verbal concept `učit-001` 'to teach', being its `ARG0` (thus `ARG0-of` relation is used); the attributes *můj* 'my' and *dějepis* 'history' are arguments of the verbal concept.

(9) *můj učitel dějepisu*
'my history teacher'

a. PDT-C (tectogrammatical layer)

```
(učitel 'teacher'
    :RSTR můj 'my'
    :RSTR dějepis 'history'
    ... )
```

b. UMR

```
(s1p1 / person
  :refer-number singular
  :ARG0-of (s1u1 / učit-001
                    'to teach'
        :ARG1 (s1d1 / dějepis
                     'history'
            :refer-num. sing.)
        :ARG2 (s1p1 / person
            :refer-num. sing.
            :refer-person 1)
        ...))
```

3.1. Data resources for deverbal nouns in UMR

To address the issues presented in the previous section in the automatic conversion of the Czech and Latin data into the UMR format, several types of data resources can be used.

PDT-C and LDT corpora. The source data used for the conversion are represented by the Prague Dependency Treebank (PDT-C))[7] (Hajič et al., 2020; Hajič et al., 2024) for Czech and the Latin Dependency Treebank (LDT)[8] (Bamman and Crane, 2006; Passarotti, 2014; Gonzalez Saavedra and Passarotti, 2014) for Latin. Both treebanks share the same dependency-based annotation scenario, which is centered on the predicate-argument structure (valency) and other deep syntactic relations. It is further enriched with semantically relevant morphological features (e.g., number and gender for nouns; tense, aspect, and modality for verbs), topic–focus articulation, and coreference annotation. More detailed comparison of the two frameworks, including an overview of automatic conversion, is presented in (Lopatková et al., 2024; Štěpánek et al., 2025a; Lopatková et al., 2025a).

PDT-Vallex lexicon. To deal with derived nouns, it is necessary to combine the source data with additional data resources. The most straightforward information for Czech nouns is provided in PDT-Vallex: for a limited number of deverbal nouns with the suffix *-ní/tí*, explanatory notes identify base verbs (but not the particular sense of the verb). Some deverbal nouns created using other word-formation processes are also included, but not in a systematic way, and their entry does not contain any information about their base verbs.

DeriNet and WFL. Furthermore, both Czech and Latin have fairly large derivational networks available: DeriNet (Olbrich et al., 2025) contains over 1 million Czech lexemes connected via word-formation links, while Word Formation Latin (WFL) (Litta et al., 2018) is a lexicon for Classical and Late Latin covering 69,682 lemmas and modeling word formation through rules represented as directed one-to-many input–output relations between lemmas. However, using these resources directly is not possible, because they do not include any information about the derived words' semantic relations to their base words (except for the identification of diminutives, female counterparts, and iterativity in DeriNet) and no information about valency. Additional steps are therefore required to link a derived noun to a particular verb-sense in the reference valency lexicon.

NomVallex lexicon. For Czech, a resource that provides information on both semantics and valency of derived nouns is the NomVallex lexicon (Kolářová et al., 2024), which contains 730 deverbal or deadjectival nouns, and deverbal, denominal, deadjectival and primary adjectives. Each word's entry comprises one or more senses, and each sense is associated with a valency frame, as well as a semantic category (such as 'action', 'abstract result of an action', 'material', 'object') and a particular verb-sense in VALLEX (Lopatková et al., 2022) from which the particular sense of the noun is derived. For example, the noun *řešení* 'solution' has the following three senses in NomVallex:

1. ∼ *seeking a satisfactory solution; discussing*
   valency: ACT(2,7,poss), PAT(2,poss)
   derived from: *řešit*-001
   semantic category: action

2. ∼ *the finding of a satisfactory solution*
   valency: ACT(2,poss,od+2), PAT(2,poss,že)
   derived from: *řešit*-001
   semantic category: abstract result of an action

3. ∼ *technical or artistic arrangement*
   valency: ACT(2,7,poss,od+2), PAT(2,poss)
   derived from: *řešit*-001
   semantic category: action / abstract result of an action

Nominals in Latin Vallex. For Latin, both Vallex 1.0 and 2.0 (and consequently Vallex4UMR) cover not only verbal predicates but also nominal and adjectival entries. However, the general issues identified for these resources (see Section 1.1) also affect the nominal domain. Most notably, frames in Vallex 1.0 lack semantic grounding, whereas those in Vallex 2.0 are not linked to LDT, except for the subset manually merged in Vallex4UMR. Unlike NomVallex for Czech, nominal entries in Latin Vallexes do not reference the verb-sense from which their particular meaning is

[7] http://hdl.handle.net/11234/1-5813
[8] https://itreebank.marginalia.it/

derived; instead, they are treated as independent lexical entries.

Smaller datasets of particular types of derivatives were also compiled to serve as material for research on Czech word-formation and can be useful in the automatic UMR annotation, namely an agent nouns dataset and a conversion dataset.

Czech agent nouns dataset. A sample of 2,828 agent nouns derived from both verbs, nouns, and adjectives using a number of different suffixes (e.g., the suffix *-tel* as in *žadatel* 'applicant' < *žádat* 'to apply', or the suffix *-ák* as in *divák* 'spectator' < *dívat se* 'to watch') was collected using the DeriNet lexicon combined with additional manual annotation.

Czech conversion dataset. A sample of pairs of suffixless nouns and their verbal counterparts, i.e., pairs of verbs and nouns where either the noun is created from the verb or the verb is created from the noun via conversion (such as *koncert* 'concert' – *koncertovat* 'to give a concert', *poprava* 'execution' – *popravit* 'execute (perf.)' / *popravovat* 'execute (imperf.)') has been compiled for Czech. This dataset is highly diverse—it contains pairs of nouns and verbs without making any decisions about the direction of conversion (deverbal vs. denominal). 50 concordances were extracted for each pair from a corpus of contemporary Czech (Křen et al., 2015) and then the concordances were manually annotated for the semantic relation between the noun and the verb in each particular occurrence, using categories such as 'action', 'result', 'agent', etc. (Ševčíková et al., 2023a; Ševčíková et al., 2023b).

3.2. Using the data resources in the automatic conversion

PDT-Vallex lexicon. To involve the annotation of eventive nouns in automatic conversion, we started by focusing on the most straightforward and systematic category of deverbal nouns: those ending with *-ní/-tí*. In total, we identified 1,690 such nouns in the PDT-C data. Using DeriNet and MorfFlex (Hajič et al., 2020), we were able to process 1,675 of these nouns and identify their base verb lexemes. These verbs are described by 2,248 valency frames (i.e., senses) in the PDT-Vallex lexicon. Almost half of them (1,062 nominal valency frames, 47 %) can be unambiguously mapped onto verbal valency frames using only participant labels; another small part can be mapped based on the morphological form of their participants.

NomVallex lexicon. This lexicon can aid the automatic conversion in cases where the participant roles are not used with the nouns in PDT-C. The process is not straightforward, because although each noun-sense in NomVallex is provided with the base verb's sense and a semantic category, there is no way of directly connecting a particular occurrence of a noun in PDT-C to a particular sense in NomVallex, as the two resources are not interlinked. However, nouns that have only a single sense in NomVallex (and can therefore be considered monosemous) can be annotated fully automatically. Out of the 730 lexemes in NomVallex, there are 91 nouns with a single sense denoting an 'action' and 34 nouns with a single sense denoting an 'abstract result of action'. In the conversion procedure, these can be either assigned an eventive concept labeled with the base verb's sense (for the former case) or such an eventive concept in combination with the `:result-of` participant relation (for the latter case).

Czech agent nouns dataset. The dataset contains lexemes along with their base word, and can therefore be used to identify nouns that should be annotated using their base verb in combination with the `:ARG0-of` inverse relation. Nouns in the dataset that are derived from parts of speech other than verbs can be disregarded, leaving 1,178 lexemes in the dataset.

An issue that needs to be addressed is that due to affix polysemy, some agent nouns can denote both an agent and an instrument; for example, the noun *nosič* 'carrier' can refer both to a person that carries something (a porter) or an instrument used for carrying something. However, this issue is easily solved in Czech because the agent nouns typically have masculine grammatical gender, and masculine nouns express animacy. The category of animacy is part of the morphological annotation in PDT-C, and it is therefore possible to automatically tell apart the non-animate (and therefore instrumental) nouns from the animate (and therefore agentive) nouns.

Once the agent noun and its base verb are identified, the particular sense of the verb that corresponds to the noun has to be found in PDT-Vallex in the next step. Out of the 1,178 agent nouns, 401 have a base verb that only has a single sense in PDT-Vallex, and this sense can therefore be assigned fully automatically. For 235 out of the agent nouns that have a base verb with multiple senses, we were able to manually identify a verb-sense that will very likely correspond to the agent noun in all of its occurrences. For example, the base verb *bruslit* 'to skate' corresponding to the noun *bruslař* 'skater' has two senses in PDT-Vallex (1. to act skillfully in some activity, 2. to skate), but the agent noun is clearly only related to the second sense of the base verb. Taken together, this means that there is

a total of 636 agent nouns that can be automatically assigned a particular verb-sense in the conversion.

Czech conversion dataset. Using this dataset presents some specific challenges. Firstly, because the dataset was compiled without making any decisions about whether the noun was converted from the verb or vice versa, it contains some nouns that are clearly not deverbal, such as *šéf* 'chief, boss' (along with its corresponding denominal verb *šéfovat* 'to be the boss'). Therefore, we only focus on the nouns that are annotated as denoting an 'action' in all corpus concordances in the sample, because although the direction of conversion is still unclear in many cases (cf. e.g. Ševčíková, 2021, for a discussion on directionality in Czech conversion), they should uncontroversially be annotated using the eventive concept in UMR. There are 257 such nouns in the dataset.

In this case, the issue connected with the Czech aspectual system (cf. Section 2.2) is prominent: 161 of these nouns have both an imperfective and a perfective corresponding verb that differs in the thematic suffix (e.g., *vyrobit* 'to produce (perf.)' and *vyrábět* 'to produce (imperf.)' for the noun *výroba* 'production'), and because the thematic suffix is not part of the noun, it is not immediately clear from the noun's form which verb should be chosen in the annotation. As we have mentioned, sometimes this is ambiguous even in a particular sentential context. We were able to manually identify the verb that will likely correspond to the noun in all its occurrences for 96 out of the 161 nouns. Furthermore, there are 90 verbs with only an imperfective verb and 6 nouns with only a perfective verb available. Therefore, it is possible to identify a single verb for 192 nouns. Out of these nouns, we were able to manually identify a particular verb-sense in PDT-Vallex that is likely to correspond to the noun in all its occurrences for 137 nouns.

Further, for these 137 nouns, we also tried to look into a procedure that would automatically label the noun's dependents as arguments of the eventive concept in case the dependents are not annotated with participant labels in PDT-C. In this procedure, we can use information about the morphological realization of arguments that is available in PDT-Vallex for 81 of the verbs; cf., e.g., the mapping for the noun *debata* 'debate':

- genitive, possessive pronoun → `ARG0`
- *o* 'about' + locative, *nad* 'over' + instrumental, *zda*, *zdali*, *jestli* 'if' → `ARG1`
- *s* 'with' + instrumental → `ARG2`

Using this mapping, the example sentence containing *debata* 'debate' given in 7 would be correctly converted into the UMR format even if it was not annotated using participant roles in PDT-C, by applying information about the dependents' forms in the morphological layer. Morphological information is not always fully deterministic, as some forms can be ambiguous as to which argument they express (cf. e.g. Kolářová et al., 2019). This is typical for instance for the genitive case, which can often refer either to the first argument or the second argument; cf. the noun *bojkot* 'boycott' in example 10, where the genitive refers to `ARG0` (the countries are the ones doing the boycotting), vs. in example 11, where the genitive refers to `ARG1` (the goods are what is being boycotted).

(10) *bojkot západních zemí*.`ARG0`
'boycott by the western countries'

(11) *bojkot zboží*.`ARG1`
'boycott of the goods'

Vallex4UMR. For Latin, Vallex4UMR can partially support automatic conversion. A subset of nouns from Vallex 1.0, which also appear in LDT, has been manually mapped to the corresponding entries in Vallex 2.0 and is therefore included in Vallex4UMR. These entries can consequently be converted with relative ease, but they only amount to 111 eventive nouns in LDT. However, beyond this subset, there is no direct way to link a specific occurrence of a noun in LDT to a particular sense in Vallex 2.0, as the two resources are not interlinked. For the remaining entries, the conversion process is further complicated by the fact that noun senses are not associated with their corresponding base verbs and semantic categories, as is the case in NomVallex for Czech. Nevertheless, nouns that are monosemous in Vallex4UMR and denote an event[9] (i.e., 464 out of the 1,857 eventive nominal lexemes in Vallex4UMR) can be annotated fully automatically. These nouns can be assigned the corresponding sense during conversion; however, the distinction between eventive and resultative interpretations cannot be made, as such information is not represented in the resource.

In summary, we have proposed how to use the available data resources for the identification of the appropriate UMR representation of derived nouns based on eventive concepts for varying numbers of nouns (Table 2). The numbers indicate nouns that can be assigned a representation with high confidence. For other nouns, certain heuristics can be applied; for example, assigning the most frequent verb-sense to polysemous nouns derived from polysemous verbs, or using the morphological information of a noun's dependents to infer

[9] In Vallex 2.0, and thus in Vallex4UMR, non-eventive nouns are also defined with senses. We identify eventive nominals by their annotation of participant roles.

correct argument labels. However, it remains unclear whether such heuristics would improve the accuracy of the automatic conversion, and whether they would be useful in creating pre-annotated data for subsequent manual correction, as they might instead introduce additional challenges for annotators. This needs to be tested in the future.

Decision trees that summarize the proposed automatic conversion are in Appendices B and C.

CZECH	
data resource	**number of nouns**
PDT-Vallex	1,062
NomVallex	135
Czech agent nouns dataset	636
Czech conversion dataset	137

LATIN	
data resource	**number of nouns**
Monosemous nouns in Vallex4UMR	464
Manually linked nouns in LDT	111

Table 2: The number of nouns that can be automatically assigned a verb-sense in the automatic conversion, for each language and resource.

4. Conclusion

In this paper, we have identified phenomena that are problematic for the annotation of deverbal nouns in Czech and Latin, both in terms of determining the appropriate annotation that would best conform to the guidelines when creating the UMR graphs manually and in terms of creating the UMR graphs via a rule-based automatic conversion from other existing data formats.

In terms of deciding on the appropriate annotation, we have discussed the issues that have to be solved in the absence of a PropBank-like lexicon that would contain sufficient coverage of nouns. As the guidelines do not offer testable criteria for identifying nouns that should be represented using an eventive concept, we suggest that deverbal nouns that clearly refer to their base verb's meaning and denote an event or an argument in the event should be annotated with an eventive concept (in combination with the appropriate inverse argument relation in the latter case). Additionally, we suggest that deverbal nouns denoting the result of an action, which is semantically very close to the eventive meaning, should be annotated with the `:result-of` relation.

When a verb corresponding to the noun has to be chosen for labeling the eventive concept, such verb may not exist in the given language, or multiple possible verbs may be available even in a specific sentential context. In case of a non-existent verb, a light-verb construction can be searched in the reference valency lexicon. When several synonymous light-verb constructions are available, they are both in principle correct and we suggest that annotators should agree on a simple rule, such as using the one that is listed first, to avoid introducing unnecessary inter-annotator disagreements into the data. Where there is no light-verb construction available in the reference valency lexicon, we suggest that a synonymous verb should be found and used instead (cf. some similar examples where a synonymous word is used instead of a multi-word expression in Bonn et al., 2023). In case there are multiple verbs corresponding to a particular noun occurrence, a systematic decision on which of them to use should be made to avoid unnecessary inter-annotator disagreements again—for the Czech aspectual pairs, we suggest preferring the imperfective verb, as it is sometimes taken to be the unmarked member of the pair (cf. e.g. Komárek et al., 1985, p. 180). The approach fully complies with the UMR ontology—all the solutions that we have proposed use existing concepts, relations and attributes.

As for the automatic conversion, the principal issue is identifying the particular sense of the deverbal noun and the particular verb-sense that corresponds to it. We have shown that there are several resources that can be used in combination with the original PDT-C and LTD data to find the corresponding verb-sense for derived nouns in a rule-based way. This is possible in cases where the noun is identified as monosemous and the base verb is either monosemous or polysemous but the noun is supposed to only relate to one of the senses in all its occurrences. Additional heuristics could be applied to other cases, such as choosing the most frequent sense of the noun and the corresponding verb for the annotation, because some of the resources do also provide information about how comparatively frequent the individual senses are. In the next step, it is necessary to test the procedures suggested for the automatic conversion on actual data and formally evaluate the result.

The procedures we suggest to apply to the Czech and Latin data in the process of manual annotation can be in principle applied to any language, although different languages may present additional problematic phenomena that we have not focused on. Extending the automatic conversion to other languages is not straightforward, as the procedure is dependent on the specific data resources that are available for a given language.

5. Acknowledgements

The work described herein has been supported by the grants *LINDAT/CLARIAH-CZ* (Project No. LM2023062) of the Ministry of Education, Youth, and Sports of the Czech Republic, GAUK No. 104924 of the Charles University, and the Charles University Research Centre program No. 24/SSH/009.

The project has been using data and tools provided by the *LINDAT/CLARIAH-CZ Research Infrastructure* (https://lindat.cz), supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2023062).

6. Ethic statement

All data used in this study are derived from publicly available language resources; therefore, no ethical approval was required and no ethical guidelines were violated.

7. Bibliographical References

David Bamman and Gregory Crane. 2006. The design and use of a Latin dependency treebank. In *Proceedings of the Fifth Workshop on Treebanks and Linguistic Theories (TLT2006)*, pages 67–78. Citeseer.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract Meaning Representation for Sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

Julia Bonn, Matthew J. Buchholz, Jayeol Chun, Andrew Cowell, William Croft, Lukas Denk, Sijia Ge, Jan Hajič, Kenneth Lai, James H. Martin, Skatje Myers, Alexis Palmer, Martha Palmer, Claire Benet Post, James Pustejovsky, Kristine Stenzel, Haibo Sun, Zdeňka Urešová, Rosa Vallejos, Jens E. L. Van Gysel, Meagan Vigus, Nianwen Xue, and Jin Zhao. 2024. Building a broad infrastructure for uniform meaning representations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2537–2547, Torino, Italia. ELRA and ICCL.

Julia Bonn, Andrew Cowell, Jan Hajič, Alexis Palmer, Martha Palmer, James Pustejovsky, Haibo Sun, Zdeňka Urešová, Shira Wein, Nianwen Xue, and Jin Zhao. 2023. UMR annotation of multiword expressions. In *Proceedings of the Fourth International Workshop on Designing Meaning Representations*, pages 99–109, Nancy, France. Association for Computational Linguistics.

William Croft. 2001. *Radical Construction Grammar: Syntactic Theory in Typological Perspective*. Oxford University Press, Oxford.

William Croft. 2022. *Morphosyntax: Constructions of the World's Languages*. Cambridge University Press, Cambridge.

Berta Gonzalez Saavedra and Marco Carlo Passarotti. 2014. Challenges in enhancing the Index Thomisticus treebank with semantic and pragmatic annotation. In *Proceedings of the Thirteenth International Workshop on Treebanks and Linguistic Theories (TLT-13)*, pages 265–270.

Jan Hajič, Eduard Bejček, Jaroslava Hlaváčová, Marie Mikulová, Milan Straka, Jan Štěpánek, and Barbora Štěpánková. 2020. Prague Dependency Treebank - Consolidated 1.0. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)*, pages 5208–5218, Marseille, France. European Language Resources Association.

Jan Hajič, Eva Fučíková, Markéta Lopatková, and Zdeňka Urešová. 2024. Mapping Czech Verbal Valency to PropBank Argument Labels. In *Proceedings of the Fifth International Workshop on Designing Meaning Representations (DMR 2024)*, pages 88–100, Torino, Italia. ELRA and ICCL.

Jan Hajič, Jarmila Panevová, Zdeňka Urešová, Alevtina Bémová, Veronika Kolářová, and Petr Pajas. 2003. PDT-VALLEX: Creating a large-coverage valency lexicon for treebank annotation. In *Proceedings of The Second Workshop on Treebanks and Linguistic Theories*, volume 9 of *Mathematical Modeling in Physics, Engineering and Cognitive Sciences*, pages 57–68, Vaxjo, Sweden. Vaxjo University Press.

Veronika Kolářová, Anna Vernerová, and Jonathan Verner. 2019. Non-systemic valency behavior of Czech deverbal nouns based on the NomVallex lexicon. *Jazykovedný časopis / Journal of Linguistics*, 70(2):424–433.

Miroslav Komárek, Jan Kořenský, Jan Petr, and Jarmila Veselková. 1985. *Mluvnice češtiny 2. Tavrosloví*. Academia, Prague.

Jerzy Kuryłowicz. 1936. Derivation lexicale et derivation syntaxique. *Bulletin de la Société de linguistique*, 32:79–92.

Markéta Lopatková, Eva Fučíková, Federica Gamba, Jan Hajič, Hana Hledíková, Marie Mikulová, Michal Novák, Jan Štěpánek, Daniel Zeman, and Šárka Zikánová. 2025a. UMR 2.0 - Czech: Release Notes. Technical Report TR-2025-74, ÚFAL MFF UK, Prague, Czechia.

Markéta Lopatková, Eva Fučíková, Federica Gamba, Jan Štěpánek, Daniel Zeman, and Šárka Zikánová. 2024. Towards a conversion of the Prague Dependency Treebank data to the Uniform Meaning Representation. In *Proceedings of the 24th Conference Information Technologies – Applications and Theory (ITAT 2024)*, pages 62–76, Košice, Slovakia. Univerzita Pavla Jozefa Šafárika v Košiciach, Košice, Slovakia, CEUR-WS.org.

Markéta Lopatková, Hana Hledíková, Jan Štěpánek, and Daniel Zeman. 2025b. From the Prague Dependency Treebank to the Uniform Meaning Representation: Gold-standard Czech UMR data and partial automatic conversion. In *Proceedings of the 25th Conference Information Technologies – Applications and Theory (ITAT 2025)*, pages 179–190, Košice, Slovakia. CEUR-WS.org.

Francesco Mambrini, Marco Passarotti, Eleonora Litta, and Giovanni Moretti. 2021. Interlinking valency frames and wordnet synsets in the LiLa knowledge base of linguistic resources for Latin. In *Further with Knowledge Graphs*, pages 16–28. IOS Press.

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

Marco Passarotti. 2014. From Syntax to Semantics. First Steps Towards Tectogrammatical Annotation of Latin. In *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)*, pages 100–109, Gothenburg, Sweden. Association for Computational Linguistics.

Marco Passarotti. 2019. The Project of the Index Thomisticus Treebank. *Digital Classical Philology*, 10:299–320.

Marco Passarotti, Francesco Mambrini, Greta Franzini, Flavio Massimiliano Cecchini, Eleonora Litta, Giovanni Moretti, Paolo Ruffolo, and Rachele Sprugnoli. 2020. Interlinking through lemmas. The lexical collection of the LiLa knowledge base of linguistic resources for Latin. *Studi e Saggi Linguistici*, 58(1):177–212.

Marco Passarotti, Berta González Saavedra, and Christophe Onambele. 2016. Latin Vallex. A Treebank-based Semantic Valency Lexicon for Latin. In *Proceedings LREC 2016*, pages 2599–2606, Portorož, Slovenia. ELRA.

Sameer Pradhan, Julia Bonn, Skatje Myers, Kathryn Conger, Tim O'Gorman, James Gung, Kristin Wright-Bettner, and Martha Palmer. 2022. PropBank comes of age—larger, smarter, and more diverse. In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pages 278–288, Seattle, Washington. ACL.

Petr Sgall, Eva Hajičová, and Jarmila Panevová. 1986. *The Meaning of the Sentence in Its Semantic and Pragmatic Aspects*. Reidel, Dordrecht.

Jan Štěpánek, Daniel Zeman, Markéta Lopatková, Federica Gamba, and Hana Hledíková. 2025a. Comparing Manual and Automatic UMRs for Czech and Latin. In *Proceedings of the Sixth International Workshop on Designing Meaning Representations (DMR 2025)*, pages 1–12, Stroudsburg, PA, USA. Association for Computational Linguistics.

Gergory Stump. 2019. Some sources of apparent gaps in derivational paradigms. *Morphology*, 29:271–292.

David H. Tuggy. 1985. The inflectional/derivational distinction. *Work Papers of the Summer Institute of Linguistics, University of North Dakota Session:*, 29:209–222.

Zdeňka Urešová, Eva Fučíková, Cristina Fernández-Alcaina, and Jan Hajič. 2025a. Linking an Event-type Ontology to Morphosyntax of the Predicate-Argument Structure. *Dictionaries: Journal of the Dictionary Society of North America*, 46(1):207–227.

Jens van Gysel, Meagan Vigus, Jayeol Chun, Kenneth Lai, Sarah Moeller, Jiarui Yao, Tim O'Gorman, James Cowell, William Croft, Chu-Ren Huang, Jan Hajič, James Martin, Stephan Oepen, Martha Palmer, James Pustejovsky, and Rosa Vallejos. 2021. Designing a uniform meaning representation for natural language processing. *KI - Künstliche Intelligenz*, 35(2):343–360.

Shira Wein and Julia Bonn. 2023. Comparing UMR and cross-lingual adaptations of AMR. In *Proceedings of the Fourth International Workshop on Designing Meaning Representations (DMR 2023)*, pages 23–33, Nancy, France. Association for Computational Linguistics.

Magda Ševčíková. 2021. Action nouns vs. nouns as bases for denominal verbs in Czech: A case study on directionality in derivation. *Word Structure*, 14(1):97–128.

Magda Ševčíková, Hana Hledíková, Lukáš Kyjánek, and Anna Staňková. 2023a. Semantics of noun/verb conversion in czech: lessons learned from corpus data annotation. *SKASE Journal of Theoretical Linguistics*, 20(4):74–92.

8. Language Resource References

Julia Bonn and Claire Bonial and Matt Buchholz and Hsiao-Jung Cheng and Alvin Chen and Ching-wen Chen and Andrew Cowell and William Croft and Lukas Denk and Ahmed Elsayed and Eva Fučíková and Federica Gamba and Carlos Gomez and Jan Hajič and Eva Hajičová and Jiří Havelka and Loden Havenmeier and Hana Hledíková and Ath Kilgore and Veronika Kolářová and Lucie Kučová and Kenneth Lai and Bin Li and Jingyi Li and Markéta Lopatková and Marie MacGregor and Marie Mikulová and Jiří Mírovský and Anna Nedoluzhko and Skatje Myers and Michal Novák and Tim O'Gorman and Petr Pajas and Alexis Palmer and Martha Palmer and Jarmila Panevová and Benét Post and James Pustejovsky and Petr Sgall and Jialin Song and Li Song and Magda Ševčíková and Jan Štěpánek and Zdeňka Urešová and Haibo Sun and Yao Sun and Rosa Vallejos Yopán and Jens Van Gysel and Meagan Vigus and Kristin Wright-Bettner and Jiawei Wu and Nianwen Xue and Dan Xing and Keer Xu and Zhixing Xu and Liulu Yue and Daniel Zeman and Jin Zhao and Šárka Zikánová and Zdeněk Žabokrtský. 2026. *Uniform Meaning Representation 2.2*. LINDAT/CLARIAH-CZ Digital Library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Jan Hajič and Eduard Bejček and Alevtina Bémová and Eva Buráňová and Eva Fučíková and Eva Hajičová and Jiří Havelka and Jaroslava Hlaváčová and Petr Homola and Pavel Ircing and Jiří Kárník and Václava Kettnerová and Natalia Klyueva and Veronika Kolářová and Lucie Kučová and Markéta Lopatková and David Mareček and Marie Mikulová and Jiří Mírovský and Anna Nedoluzhko and Michal Novák and Petr Pajas and Jarmila Panevová and Nino Peterek and Lucie Poláková and Martin Popel and Jan Popelka and Jan Romportl and Magdaléna Rysová and Jiří Semecký and Petr Sgall and Johanka Spoustová and Milan Straka and Pavel Straňák and Pavlína Synková and Magda Ševčíková and Jana Šindlerová and Jan Štěpánek and Barbora Štěpánková and Josef Toman and Zdeňka Urešová and Barbora Vidová Hladká and Daniel Zeman and Šárka Zikánová and Zdeněk Žabokrtský. 2024. *Prague Dependency Treebank - Consolidated 2.0 (PDT-C 2.0)*. LINDAT/CLARIAH-CZ Digital Library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Jan Hajič and Jaroslava Hlaváčová and Marie Mikulová and Milan Straka and Barbora Štěpánková. 2020. *MorfFlex CZ 2.0*. LINDAT/CLARIAH-CZ Digital Library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Veronika Kolářová and Václava Kettnerová and Jana Klímová and Jiří Mírovský and Anna Vernerová. 2024. *NomVallex 2.5*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).

Michal Křen and Václav Cvrček and Tomáš Čapka and Anna Čermáková and Milena Hnátková and Lucie Chlumská and Tomáš Jelínek and Dominika Kováříková and Vladimír Petkevič and Pavel Procházka and Hana Skoumalová and Michal Škrabal and Petr Truneček and Pavel Vondřička and Adrian Jan Zasina. 2015. *SYN2015: A Representative Corpus of Written Czech*. Prague, Institute of the Czech National Corpus, Faculty of Arts, Charles University; http://www.korpus.cz.

Litta, Eleonora and Passarotti, Marco and Culy, Chris. 2018. *Morphology Beyond Inflection. Building a Word Formation-Based Lexicon for Latin*. Cambridge Scholars Publishing, Newcastle upon Tyne.

Markéta Lopatková and Václava Kettnerová and Jiří Mírovský and Anna Vernerová and Eduard Bejček and Zdeněk Žabokrtský. 2022. *VALLEX 4.5*. LINDAT/CLARIAH-CZ Digital Library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Michal Olbrich and Viktória Brezinová and Šárka Dohnalová and Vojtěch John and Lukáš Kyjánek and Aleš Papáček and Emil Svoboda and Magda Ševčíková and Jonáš Vidra and Zdeněk Žabokrtský. 2025. *DeriNet 2.3*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).

Ševčíková, Magda and Kyjánek, Lukáš and Hledíková, Hana and Staňková, Anna. 2023b. *Semantic annotation of noun/verb conversion in Czech*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).

Jan Štěpánek and Markéta Lopatková and Daniel Zeman and Federica Gamba and Hana Hledíková and Eva Fučíková and Michal Novák and Šárka Zikánová and Eva Hajičová and Jiří Havelka and Veronika Kolářová and Lucie Kučová and Marie Mikulová and Jiří Mírovský and Anna Nedoluzhko and Petr Pajas and Jarmila Panevová and Petr Sgall and Magda Ševčíková and Zdeňka Urešová and Zdeněk Žabokrtský and Jan Hajič. 2025b. *Uniform Meaning Representation 2.1 (Czech and Latin)*. LINDAT/CLARIAH-CZ Digital Library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Zdeňka Urešová and Eva Fučíková and Jan Hajič and Veronika Kolářová and Cristina Fernández Alcaina and Peter Bourgonje and Eva Hajičová and Georg Rehm and Kateřina Rysová and Karolina Zaczynska. 2025b. *SynSemClass 5.5*. LINDAT/CLARIAH-CZ Digital Library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Zdeňka Urešová and Alevtina Bémová and Eva Fučíková and Jan Hajič and Veronika Kolářová and Marie Mikulová and Petr Pajas and Jarmila Panevová and Jan Štěpánek. 2021. *PDT-Vallex: Czech Valency lexicon linked to treebanks 4.0 (PDT-Vallex 4.0)*. LINDAT/CLARIAH-CZ Digital Library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Appendix A: Decision Tree for Manual Annotation (Czech and Latin)

For each noun decide whether it denotes / relates to an event!

Does the **noun itself denote** an event?

YES: Represent it as an eventive concept. (Add argument relations, see below!)
e.g., *běhání* 'running', *příchod* 'arrival', *akvizice* 'acquiring'

- Does **a single corresponding verb** exist in the valency lexicon?
  - YES: Use the verb's lemma + particular sense in the valency lexicon.
    e.g., *běhání* --> běhat-002
  - NO: Do **multiple corresponding verbs** exist in the valency lexicon?
    - YES: Is it an aspectual pair?
      - YES: Use the imperfective verb.
        e.g., *příchod* --> přicházet-008
      - NO: Use the one listed first in the valency lexicon.
    - NO: (= no corresponding verb)
      - Does **a single LVC exist** in the v. lexicon with the given noun?
        - YES: Use the light-verb's lemma + the noun's word-form.
          e.g., *akvizice* 'acquiring' --> provést-akvizici-004
        - NO: Do **multiple LVCs** exist?
          - YES: Use the one listed first
          - NO: (= no LVC exists)
            Use a synonymous verb.
            e.g., *publicita* 'publicity' --> propagovat-001

NO: (= the noun does not denote event)

- Does the noun denote **an argument** of an event?
  - YES: Represent it as an abstract concept
    with the event attached using the given :ARGx-of relation.
    e.g., *učitel* 'teacher' --> (p / person :ARG0-of (u / učit-001 'teach'))
    e.g., *nabídka* '(an) offer' --> (t / thing :ARG1-of (n / nabídnout-001 '(to) offer')
  - NO: Does the noun denote **an abstract result** of an event?
    - YES: Represent it as an abstract concept (typically thing)
      with the event attached using the :result-of relation.
      e.g., *informace* 'information' (t / thing :result-of (i / informovat-001 'inform'))
      e.g., *plán* '(a) plan' (t / thing :result-of (p / plánovat-001 '(to) plan'))
    - NO: Represent as entity concept (labeled with noun's lemma)

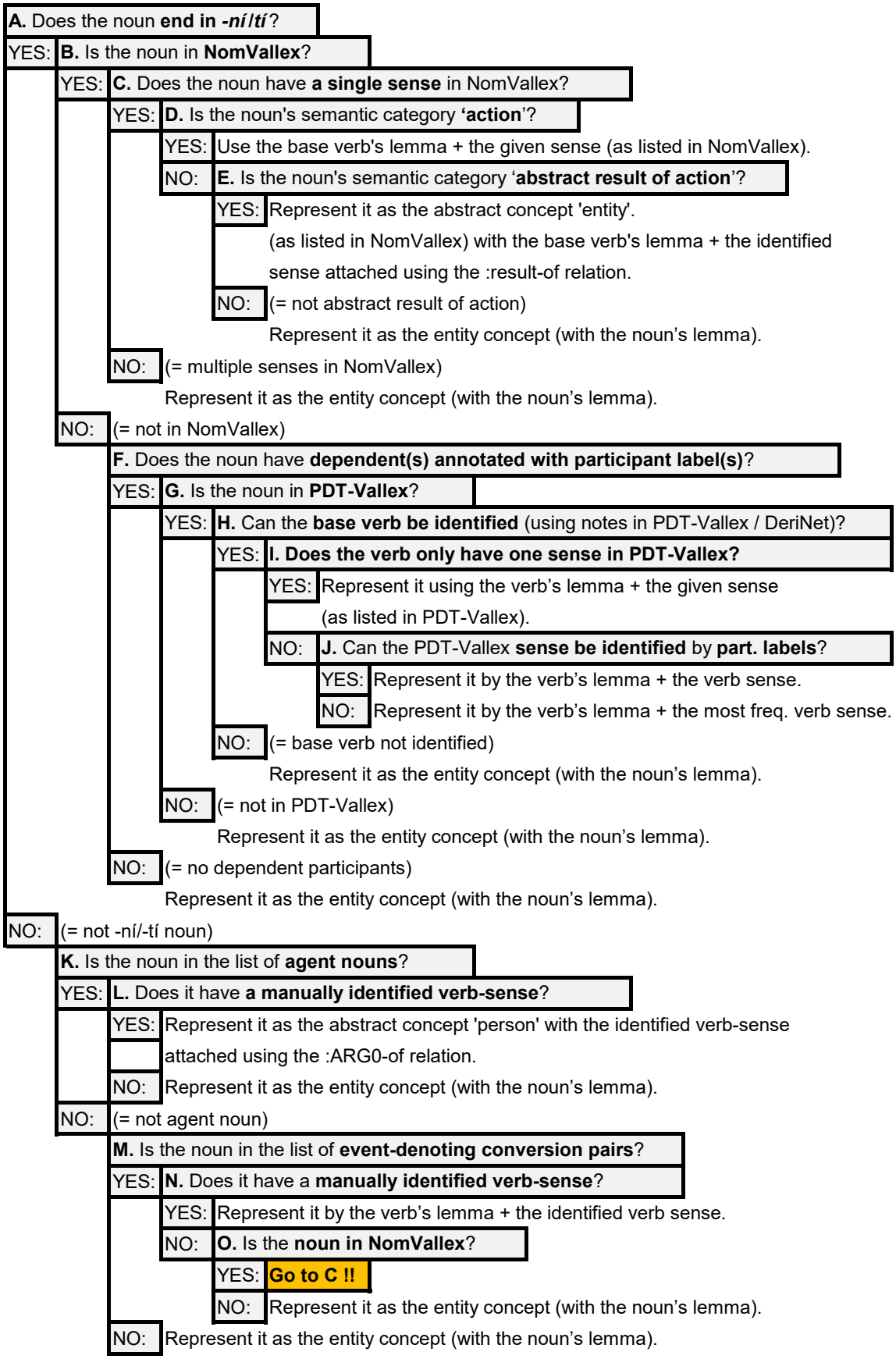
If the eventive concept is chosen: add argument relations!

- Does the dependent correspond to **an argument of the eventive** concept?
  - YES: Represent the dependent's concept as an ARG relation of the eventive concept.
  - NO: Represent the dependent's concept as a non-argument role relation
    of the eventive concept.

Add any other arguments listed in the valency lexicon as obligatory as (unspecified) ARG relations!

Appendix B: Decision Tree for Automatic Conversion (Czech)

For each noun decide how to represent it!



Appendix C: Decision Tree for Automatic Conversion (Latin)

For each noun decide how to represent it!

A. Is the noun mapped onto corresponding **entry in Vallex 2.0**?	
YES:	Represent it by the noun's lemma + the given noun sense.
NO:	**B.** Is it a **monosemous eventive noun** in **Vallex4UMR**?
	YES: Represent it by the noun's lemma + the given noun sense.
	NO: Represent it as the entity concept (with the noun's lemma).

If the eventive concept is chosen: add argument relations!

A. Does the noun has dependents annotated with participant labels?	
YES:	Use the same mapping algorithm that is used for verbs.
NO:	**B.** Is there an **unambiguous mapping** based on morphological form available?
	YES: Use morphological layer to map forms onto arguments.
	NO: Leave arguments unmapped.