Linking Two Lexical Resources: VALLEX and MorfFlex Lexicons

Markéta Lopatková*, Jaroslava Hlaváčová and Jiří Mírovský

Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics, Malostranské nám. 25, Prague, 118 00, Czech Republic

Abstract

The article focuses on two different lexicons providing complementary information: MorfFlex covering general Czech morphology, and VALLEX giving information on the syntax and semantics of Czech verbs. We discuss different designs of these lexicons, concentrating primarily on variants and homographs in the Czech vocabulary. Within the project, we have verified the theoretical approaches and harmonized the treatment of variants in both lexicons, adopting the clear morphologically based criteria from MorfFlex for distinguishing variants in VALLEX. The two updated lexicons, MorfFlex and VALLEX, with interlinked records represent the project's main outcome.

Keywords

morphology, valency, variants, homographs, linking resources

1. Motivation

Language resources represent a crucial prerequisite for any natural language processing (NLP) task – and this is true not only in the "pre-AI-chat-applications world", but even more so in the world using large language models and their AI processing, for which a large *quantity of data* is crucial. Less obvious is the usability of *high-quality* (but necessarily very limited) language resources prepared manually, i.e., data with added expert information based on (not only) linguistic erudition.

Despite these fundamental doubts concerning the usefulness of human-developed data in future NLP, we do not want to abandon and throw away these high-quality language resources yet. Still, we want to maintain them and connect different types of data as much as possible. Moreover, such high-quality data, offering deep linguistic insight, represent an essential resource for further theoretical and formal linguistics research.

Here we focus on two language resources – two different lexicons providing complementary information: MorfFlex covering general Czech *morphology* and VALLEX giving information on *syntax* and *meaning* of Czech verbs. Our goal is to merge these two resources – simply by interlinking them.

ITAT 2023: Information Technologies – Applications and Theory, 2023
lopatkova@ufal.mff.cuni.cz (M. Lopatková);
hlavacova@ufal.mff.cuni.cz (J. Hlaváčová);
mirovsky@ufal.mff.cuni.cz (J. Mírovský)
0000-0002-3833-9611 (M. Lopatková); 0000-0001-6506-6797 (J. Hlaváčová); 0000-0003-2741-1347 (J. Mírovský)

2. MorfFlex

Morphological dictionaries serve as inventories of all wordforms of a natural language, and as such, they represent essential language resources.

MorfFlex [1, 2], the morphological dictionary of the Czech language, has two main purposes:

- analysis of wordforms
- generation of wordforms

As for the first purpose, MorfFlex serves as the basis for morphological taggers, which assign a basic wordform (lemma) and a set of its morphological properties (in the form of a morphological tag) to every Czech wordform. In NLP tasks, this helps to reduce data sparsity, as it allows automatic tools to work just with lemmas (instead of individual wordforms), or just with morphological tags. Lemmatization and tagging also allow machines (and human users as well) to create queries for effective searching in language corpora.

From the other side, the morphological dictionary contains all the necessary information for generating wordforms based on their lemma and morphological tag. It is used in various tools of NLP, for instance, machine translation.

MorfFlex is maintained as a set of triplets
`<wordform, lemma, tag>`.

The lemma is a basic (representative) wordform, which usually serves as the key word in dictionaries. In MorfFlex, it can be accompanied by a brief semantic note which serves only human editors. It means that MorfFlex does not contain any information concerning syntactic or semantic properties of words in a form that could be used in automatic tools.

For example, the following triplet
`<pískal, pískat, VpYS----R-AAI-->`

is the MorfFlex entry belonging to the wordform *pískal* '(he) whistled'. Given the wordform (the first item from the triplet), MorfFlex provides its lemma and its morphological description in the form of a morphological tag[1] (compare the first purpose from above). On the other hand, given a lemma and a morphological tag (the second and the third item), the particular wordform can be generated.

Putting it differently, MorfFlex associates individual lemmas with the complete set of their wordforms, i.e., with their whole (morphological) paradigms. In other words, MorfFlex provides complete characteristics of the formal part of Czech lexemes (as described in Section 3).

MorfFlex contains more than 1 million lemmas, 56,748 of which are verbs.

One important rule applies in MorfFlex, referred to as the *golden rule of morphology* in [4]: Any pair <lemma, tag> can be associated with only a single wordform, not more. This requirement guarantees that a unique wordform is derived based on the lemma and tag. To put it differently, it cannot happen that a particular lemma together with a particular tag are attached to more than one wordform. This rule is essential for the unambiguous description of variants. Namely, every variant has its unique description – a unique record in MorfFlex. As an illustration of the principle, compare the two variants of the imperative of the verb *plavat* 'to swim', with their tags differing in the last position:[2]

```
<plav, plavat, Vi-S---2--A-I-->
<plavej, plavat, Vi-S---2--A-I-1>
```

A detailed description of MorfFlex and the adopted principles can be found in [2].

As described above, MorfFlex contains complete morphological characteristics of (all) Czech wordforms, including their grouping into paradigms represented by a lemma. As such, it is a highly valuable source of information on morphology for different types of lexicons like VALLEX.

[1] The morphological tag is a label indicating the morphological features of a wordform. In MorfFlex, it is structured as a string of 15 positions, as described in [3]. Here the first position indicates part-of-speech (N for noun, A for adjective, V for verb, etc.); the following positions are most relevant for verbs:
3 - gender (e.g., F feminine, M masculine animate),
4 - number (P plural, S singular),
8 - person (e.g., 1, 2),
9 - tense (e.g., P present, R past, F future),
12 - voice (A active, P passive) and
13 - aspect (I imperfective, P perfective, B biaspectual).
The last position distinguishes variants, see also sect. 4.3.

[2] These two records are an example of inflectional variants, see section 4.3 for more details.

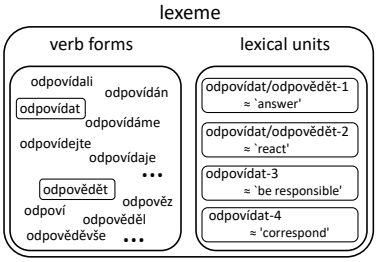

Figure 1: Lexeme, lexical forms, and lexical units in VALLEX (lemmas as the representative verb forms in boxes).

3. VALLEX

One of the key tasks linguists and NLP specialists intensively study is the possibility of representing natural language semantics. On the linguistic side, in the last 20 years, a lot of effort has been devoted to dictionaries of semantic propositions capturing predicate-argument relations or, in other words, the valency potential of predicates: verbs, nouns, adjectives, and adverbs. Let us mention esp. the FrameNet project [5], PropBank [6], VerbNet [7], OntoNotes [8].

VALLEX,[3] the Valency Lexicon of Czech Verbs [9, 10], is a collection of linguistically annotated data and documentation. It provides a formal, machine-readable description of *verbal predicates* (the term verbal predicate refers here to the meaning counterpart of a "morphological" verb), focusing on the valency properties of Czech verbs and their additional syntactic and semantic characteristics. The lexicon covers common senses of the most frequent Czech verbs: in total, it comprises 11,132 verb senses of almost 4,700 verbs, i.e., more than 6,850 verb lexical units (counting perfective and imperfective verbs as forming a single lexeme, see below). If iterative verbs are also counted, the lexicon covers 5,098 verb lemmas. The lexicon design stresses versatile usability both for a human user and for the automatic processing of Czech.

As for the VALLEX formal structure, its basic building blocks correspond to individual lexemes [11, 12] – as sketched in Figure 1, a *lexeme* is understood as a two-fold abstract entity associating:

- a set of all relevant verb forms (the whole verb morphological paradigms) represented by a set of their *lemmas*, and
- a set of *lexical units* corresponding to individual meanings, i.e., "complexes with (relatively) stable, discrete semantic properties", according to [13].

[3] https://ufal.mff.cuni.cz/vallex.
The data can be downloaded from the LINDAT/CLARIAH-CZ Repository http://hdl.handle.net/11234/1-4756.

In VALLEX, following the tradition of the Functional Generative Description and the valency theory developed within this approach [14, 15], several "morphological" verbs are typically subsumed under a single lexeme, namely:

(i) morphological variants (in a strict sense as described below in Section 4.3), marked off by the slash symbol, e.g., *vystříhávat / vystřihávat*$^{\text{impf}}$ 'to cut out',

(ii) perfective and imperfective counterparts, distinguished by their aspect in the subscript, e.g., *dávat*$^{\text{impf}}$–*dát*$^{\text{pf}}$ 'to give', and

(iiii) other verbs traditionally considered as the same predicates, e.g., *spolknout*$^{\text{pf1}}$ – *spolykat*$^{\text{pf2}}$ 'to swallow', distinguished by digits following the aspect value.[4]

In these cases, all the forms are conceived as forms of a single verbal predicate at the syntactic and semantic layer of the language description. As exemplified in Figure 2, the predicate *odpovídat*$^{\text{impf}}$ – *odpovědět*$^{\text{pf}}$ with the meaning 'to answer' is represented by two lemmas (differing in aspect). Similarly, two lemma variants *stéci/stéct*$^{\text{pf}}$ 'to flow down', are subsumed together with their imperfective counterpart *stékat*$^{\text{impf}}$ under the single lexeme *stékat*$^{\text{impf}}$ – *stéci/stéct*$^{\text{pf}}$ (Figure 3). However, individual lexical units can be associated with just a subset of lemmas, as exemplified in Figure 1, with four lexical units for the imperfective *odpovídat*$^{\text{impf}}$ and just two lexical units for the perfective *odpovědět*$^{\text{impf}}$.

VALLEX provides rich syntactic and semantic information for each lexical unit. This information includes, first of all, its valency characteristics in the form of a valency frame; supplementary information such as, e.g., types of applicable diatheses or the possibility to express reflexivity and reciprocity is available, as illustrated in Figure 2. On the other side, the information concerning relevant forms is limited to the list of relevant lemmas, i.e., infinitive forms of verbs representing the whole morphological paradigms, and their aspect (as it is syntactically and semantically relevant).

Interlinking VALLEX with MorfFlex is thus a natural way how to add detailed morphological information to VALLEX.

4. Identification of Corresponding Entries

Several differences originating from the different designs of the two lexicons (as sketched above) had to be solved. First, two groups of lemmas are more or less systematically excluded from VALLEX but covered in MorfFlex:

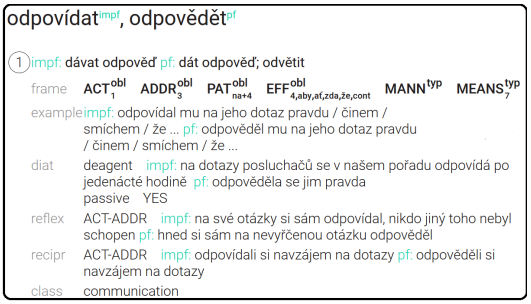

Figure 2: Lexical unit in VALLEX (predicate *odpovídat* – *odpovědět* 'to answer').

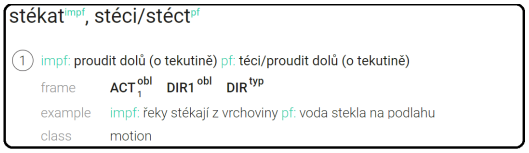

Figure 3: Lexical unit in VALLEX (predicate *stékat* – *stéci/stéct* 'to answer').

iteratives as, e.g., *běhávat* 'to run (repeatedly)' (only sporadically covered by VALLEX), and non-standard variants as, e.g., *voprášit* as a variant of the standard *oprášit* 'to dust' (ignored in VALLEX at all). In this case, it is not a mistake, VALLEX ignores them as a rule, so we simply do not process them.

Another difference concerns the forms of lemmas stored both in VALLEX and MorfFlex and related to reflexives (Sect. 4.1) and to asymmetries in treating homographs (Sect. 4.2) and variants (Sect. 4.3).

4.1. Reflexive morphemes

MorfFlex. In MorfFlex, every lemma corresponds to a "morphological" word, i.e., to a sequence of letters delimited by spaces.[5]

For example, *odpovídat* 'to answer' represents a single "morphological" word and a "semantic" word as well. However, there are "semantic" words in Czech consisting of two strings of letters, as, e.g., *bát se* 'to fear' or *povšimnout si* 'to notice'– in MorfFlex, such cases split into two entries, one for the verb lemma (*bát* or *povšimnout*) and one for the reflexive (*se* or *si*), though they cannot be used without reflexives in well-formed sentences (**Karkulka* ***bála*** *vlka.*, **Karkulka* ***oblíbila*** *vlka.*).[6]

[4] See also footnote 7 in Sect. 4.3.

[5] The same principle is applied to all wordforms, i.e., analytical forms of verbs are not covered as units in MorfFlex;e.g., the analytical form *budou odpovídat* '(they) will answer' is covered by two triplets, namely
`<budou, být, VB-P---3F-AAI-->` and
`<odpovídat, odpovídat, Vf--------A-I-->`.

[6] Unless the reflexive is elided from the surface sentence, as, e.g.,

VALLEX. In VALLEX, on the other hand, "semantic" words are treated as integral units. Thus, the pairs *bát se* 'to fear' or *povšimnout si* 'to notice' form indivisible units considered as verb lemmas. Similarly, the pair *dít se* 'to happen' is treated here as a single unit (in addition to the verb *dít* 'to tell', see the following section 4.2 dealing with homographs). In total, VALLEX covers 204 such lemmas (called reflexiva tantum).

Further, for some verbs, both non-reflexive and reflexive counterparts appear in VALLEX (they are interlinked, as their meanings are related but their syntactic patterns differ), as, e.g., *bavit (někoho)* 'to amuse (sb)' and *bavit se (s někým)* 'to talk (with sb)'. Such pairs share their morphological paradigms (differing only in the absence/presence of the reflexive), thus they are represented as a single non-reflexive verb in MorfFlex. There are 1,490 lemmas in VALLEX that appear both without and with the reflexive.

In addition, VALLEX distinguishes verbs with optional reflexives, i.e., verbs that appear both without the reflexive and with the reflexive, despite having the same syntactic pattern and the same meaning (202 verb lemmas), as. e.g., *mrknout (se)* 'to glance' (compare the following corpus example and its modification, *Počkejte minutku,* ***mrknu se****, kde by mohly být.* (SYN v10) – *Počkejte minutku,* ***mrknu****, kde by mohly být.*).

Solution. When interlinking the two lexicons, we ignore the potential reflexives *se, si*, both obligatory and optional ones. As a consequence, both non-reflexive and reflexive counterparts in VALLEX are mapped onto a single lemma in MorfLex, as, e.g., *bít* 'to beat' and *bít se* 'to fight' are mapped onto the MorfFlex lemma *bít*.

4.2. Homographs

The second asymmetry between MorfFlex and VALLEX concerns ambiguous verbs, i.e., verbs with the same lemma but (substantially) differing in their meaning, and thus considered independent units from the semantic point of view.

MorfFlex. MorfFlex does not consider meaning, only morphology. Consequently, two words with (even totally) different meanings do not have separate records unless their morphological features differ. In other words, if the paradigms of the two words are identical, they are not distinguished in MorfFlex.

For example, the verb *dít* has two different meanings, also varying in the aspect value, each of them with its unique paradigm (compare the forms of the 3rd person singular, present tense: *dí* '(he) tells' vs. *děje se* '(it) happens'), so it is necessary to have two records to distinguish them. Thus the lemma *dít-1_(dít_se)* 'to happen' or 'to occur' (imperfective) differs from the lemma *dít-2_(říkat)* 'to tell' (biaspectual).

On the other hand, the verb *topit*, despite having two clearly different meanings, namely 'to produce heat' and 'to drown', is represented by a single lemma, as in both meanings, the verb *topit* has an identical set of wordforms with the same morphological tags.

This simple and consistent criterion based on the morphological paradigm works well at the morphological level. However, it is inappropriate for VALLEX, where we accent syntax and semantics over morphology.

VALLEX. In VALLEX, verbs with different meanings but clearly etymologically connected are treated as separate lexical units within a single lexeme, as illustrated, e.g., by the predicate *odpovídat*$^{\text{impf}}$ – *odpovědět*$^{\text{pf}}$ in Figure 1.

In the same way, VALLEX treats the imperfective verb *nakupovat* 'to buy' as a counterpart to the perfective verb *nakoupit* 'to buy' in the single lexeme *nakupovat*$^{\text{impf}}$ – *nakoupit*$^{\text{pf}}$, as these two verbs share the same valency patterns. However, the lemma *nakupovat* can also be treated as the aspectual counterpart to the verb *nakupit* 'to heap'. Thus, it is necessary to distinguish them as homographs in VALLEX, *nakupovat*$_{I}$$^{\text{impf}}$ – *nakoupit*$^{\text{pf}}$ 'to buy' and *nakupovat*$_{II}$$^{\text{impf}}$ – *nakupit*$^{\text{pf}}$ 'to heap', even though they are subsumed under the single lemma *nakupovat* in MorfFlex as they represent the single "morphological" verb (having the same paradigm).

Distinguishing homographs as sketched above allows us to follow the theoretical assumptions – adopted from the Functional Generative Description (see Sect. 3) – that aspectual counterparts form a single lexeme. Thus, VALLEX can cope with asymmetrical cases where a single "morphological" verb characterizes (a formal part of) two or more lexemes.

However, in contrast to the clear technical criterion adopted in MorfFlex, the semantically (or etymologically) based criterion used in VALLEX is somewhat blurry. Consequently, different lexicons differ in their treatment of homographs since experts from time to time disagree in their interpretations (or their analysis shifts in time as the particular meaning becomes more independent in everyday practice). For example, the comprehensive *Slovník spisovného jazyka českého (SSJČ)* from the sixties [17] treats the verb *hradit* in a single entry in all its meanings (incl. 'to fence; to enclose'; and 'to cover; to reimburse') while its more recent (and substantially smaller) successor *Slovník spisovné češtiny pro školu a veřejnost (SSČ)* [18] distinguishes two homographs, *hradit*$_{I}$ 'to fence; to enclose' and *hradit*$_{II}$ 'to cover; to reimburse' (VALLEX fol-

in a dialog („*Ty* ***se bojíš*** *samoty? "* „***Bojím***.*"* ' "Are you afraid of being alone?" "−−−Yes, I am (afraid)." '), or shared in more complex usages (*Pepovi* ***si*** *o peníze* ***bála říct***. 'She was afraid to ask Pepa for the money.', where the only reflexive *si* appears instead of two ones, *se* belonging to *bát* and *si* belonging to *říct*, cf. haplology [16]).

lows the latter solution, distinguishing two homographs). The predicate *odpovídat*$^{\text{impf}}$ – *odpovědět*$^{\text{pf}}$, see Figure 1, serves as another example of fuzzy boundaries between individual lexemes – all its meanings are traditionally considered as belonging to the same lexeme, despite their substantial distance (at least from the synchronous point of view).

VALLEX (in its latest public version 4.5) distinguishes 245 homographs formed out of 122 lemmas (i.e., when ignoring the homograph marker and the possible reflexive).

Solution. When interlinking the two lexicons, we ignore the homograph marker in VALLEX and distinguish different lemmas only if they are also distinguished in MorfFlex. Consequently, two or more (homographic) verbs in VALLEX may be mapped onto one or more lemmas in MorfFlex.

In most cases (for 35 lemmas), the information on the aspect of the verb (in the morphological tag) makes it possible to detect appropriate mapping automatically. Rare cases (5 lemmas) where automatic mapping is not possible are checked and resolved manually. For example, VALLEX contains the homographic lemma *stát* and distinguishes three verbs: $stát_I^{\text{impf}}$ 'to cost', $stát_{II}^{\text{impf}}$ 'to stand; to be located', and $stát\ se_I^{\text{pf}}$ 'to happen'. On the MorfFlex side, three verbs with the samme lemma appear as well, two imperfective verbs, *stát-3_^*(stojím_stojíš) 'to stand' and *stát-5_^*(sníh) 'to melt', and one perfective, *stát-2_^*(stanu_staneš) 'to happen'. While there is a single pair of perfective verbs on both sides, the mapping is unproblematic. However, both imperfective verbs from VALLEX correspond to the imperfective verb *stát-3_^*(stojím_stojíš) (and *stát-5_^*(sníh) 'to melt' is not covered by VALLEX at all). Here manual intervention is necessary.

4.3. Morphological variants

The third asymmetry between MorfFlex and VALLEX concerns morphological variants with identical syntactic and semantic characteristics. In VALLEX, several verb lemmas with the same meaning and the same syntactic and semantic characteristics are typically grouped into one entry, as illustrated above.[7] In MorfFlex, the approach to variants is different. Variants are only described from the morphological or orthographic point of view, regardless of syntax or semantics.

MorfFlex. Since the latest version, MorfFlex CZ 2.0 [2], two types of variants are recognized and consistently marked in the dictionary.

(i) Global variants. Global variants (also called full-paradigm variants) are those variants that relate to all wordforms of a paradigm, and always in the same way, e.g., *vystříhávat* and *vystřihávat* ('to cut out', imperfective) – the whole paradigms of the verbs differ in the alternation *-í-* vs. *-i-* in the root.

Each of the two (or more) global variants has its own lemma with a complete paradigm. One of the variant lemmas is proclaimed a basic one, and the other contains a link to the basic one. In such a way, we have the variants interconnected. In the previous example, *vystříhávat* is the basic lemma. The lemma *vystřihávat* contains the link to *vystříhávat* (*vystřihávat -> vystříhávat*).

(ii) Inflectional variants. Inflectional variants (also called wordform variants) are those variants that relate only to some wordforms of a paradigm. In that case, (i) the two (or more) variants have the same lemma and (ii) all the values of all morphological categories are identical. For example, the wordforms *kopá* and *kope* '(he/she) digs' belong to the paradigm of the verb *kopat* 'to dig'; their morphological features are identical (3rd person singular, present tense). As this variant manifests just in this pair, they are considered inflectional variants (not global). The distinction is expressed through numbers in their morphological tags (at the very last position). The triplets, for example, are:

```
<kopá, kopat, VB-S---3P-AAI-->
<kope, kopat, VB-S---3P-AAI-1>
```

Inflectional variants of infinitives. There is an inflectional variant concerning the great majority of verbs that manifests itself in their lemmas. It is the common ending *-t* variant vs. the archaic ending *-ti*. The infinitive *kopat* 'to dig' of the previous example has the inflectional variant *kopati*. There is no need to artificially create two paradigms (*kopat*, *kopati*) differing only in the infinitive, they are both subsumed under the lemma *kopat*.

```
<kopat, kopat, Vf--------A-I-->
<kopati, kopat, Vf--------A-I-2>
```

Further, there are several verbs with another pair of infinitive ending variants, namely *-ci* vs. *-ct*, for instance, *říci* vs. *říct* 'to tell'. Again, those variants are inflectional since they relate to the infinitives only (subsumed under the lemma ending with *-ci*).

We should mention one more type of inflectional variants of infinitives, namely the endings *-it* vs. *-et / -ět*, as manifested with the verb *muset* 'must' and with the verb *chraptět* 'to rasp'.

```
<muset, muset, Vf--------A-I-->
<musit, muset, Vf--------A-I-1>
<chraptět, chraptět, Vf--------A-I-->
<chraptit, chraptět, Vf--------A-I-1>
```

In this case, not only the infinitives but also the wordforms of the past tense show the same difference, e.g.,

[7] Traditional Czech lexicography does not provide a testable criterion for distinguishing variants; thus, the concept of variants applied in the older VALLEX versions (3 and 4) is broader than in MorfFlex.

-il vs. *-el / -ěl*. However, the rest of the wordforms are identical, so according to the definition, the variants are inflectional.

VALLEX. As VALLEX registers just lemmas (as the representative forms of the whole paradigm), only those variants that affect lemmas are relevant for the mapping. All such variants should be mapped onto MorfFlex. In particular, VALLEX explicitly keeps all lemmas representing global variants and all lemmas representing inflectional variants where the lemma is hit. The lemma variant *-t, -ti* represents the only systematic exception, where only the first forms are present in VALLEX.

VALLEX 4.5 (the latest released version) contains 134 groups of lemma variants (ignoring possible homograph markers and reflexives). Mostly, there are two variants for a lemma; in 5 cases, there are three variants (e.g., *svléci/svléct/svlíct* 'to undress'.

Solution. As a consequence, **global lemma variants** in VALLEX are mapped onto all (interconnected) MorfFlex lemmas, for example:
VALLEX: *vystřihávat/vystříhávat* 'to cut out'
-> MorfFlex: *vystřihávat*, *vystříhávat*;
VALLEX: *oddechnout/oddychnout/oddýchnout* 'to breathe out; to rest'
-> MorfFlex: *oddechnout*, *oddychnout*, *oddýchnout*.

On the other hand, **inflectional variants affecting lemmas** (listed in VALLEX as well)[8]are mapped onto the basic lemmas in MorfFlex only, for example:
VALLEX: *říci / říct* 'to tell' -> MorfFlex: *říci*
VALLEX: *muset / musit* 'must' -> MorfFlex: *muset*

5. Linking the lexicons

Compiling the list of records for interlinking. The primary and obvious task was to compile a list of lemmas covered by both lexicons. First, we collected the set of lemma-aspect pairs from VALLEX (typically more lemmas from a single lexeme), ignoring possible reflexives (Sect. 4.1) and homograph markers (Sect. 4.2). This list of 3,635 lemma–aspect pairs served as the initial repertoire of lemmas that should be processed.

Second, we matched them with the appropriate records in MorfFlex. On the way, we found several verbal lemmas not covered by MorfFlex by mistake (e.g., *mlet* as an inflectional variant of *mlít* 'to melt'). The relevant ones were added to MorfFlex (8 lemmas in total), and the rest of them (5 archaic lemmas, as *pékat* 'to used to bake') were removed from VALLEX.

[8]Naturally, inflectional variants not affecting lemma are disregarded in VALLEX.

Third, we detected additional lemmas marked in MorfFlex as variants of the already matched lemmas (Sect. 4.3) and added them to the list (12 lemmas in total, as, e.g., colloquial *oblíct* as a variant of *obléci* 'to dress'). They were added to VALLEX as well.

Interlinking the records. The lemma candidates selected for interlinking were automatically checked – the unambiguous MorfFlex lemmas with the same aspect value as the VALLEX ones were automatically added to the new VALLEX attribute `-morfflex` assigned to each lexical unit.

Aspect. The lemma pairs differing in aspect were manually checked (35 cases). Typically, the variance was caused by the diverse classification of these verbs in the source Czech lexicons (as SSJČ or SSČ mentioned in Sect. 4.2) and in their corpora usage reflecting current Czech and its development (as, e.g., *dovést*$_{II}$ 'to be able', which obviously moves from imperfective to perfective). In these cases, the VALLEX and MorfFlex aspect values were harmonized and the pairs of records were interlinked.

Homographs. There were also several cases of ambiguous mapping detected by the automatic procedure, i.e., cases of two MorfFlex lemma candidates with the same aspect value identified as possible counterparts of a particular record in VALLEX. These few cases had to be resolved manually (5 lemmas).

As a by-product, several cases of inappropriate homograph splitting were corrected (namely, 4 archaic lemmas were removed from VALLEX based on the MorfFlex evidence).

Variants. The most significant part of the changes concerned the harmonization of variants due to the strictly morphology-based criterion used in MorfFlex for distinguishing variants. On the MorfFlex side, four inflectional variants (e.g., *dožnout / dožít* 'to finish mowing') and 14 global variants (e.g., *nadechnout / nadýchnout* 'to inhale') were added. As for VALLEX, 47 variants (as detected in previous releases) were separated as non-variants. Nevertheless, the particular lemmas were kept in the same lexical units but marked as non-variants in the updated data (as, e.g., *plavat / plovat*$^{\text{impf}}$ was replaced by *plavat*$^{\text{impf1}}$ – *plovat*$^{\text{impf2}}$). On the other hand, two pairs of verbs were connected as variants in VALLEX based on MorfFlex (e.g., *utvářet / utvářit*$^{\text{impf}}$ 'to create'). Further, as mentioned earlier, several new variants were added, covering mainly colloquial Czech (including 3 new variants with homograph markers).

The updated VALLEX data cover 108 variant groups (ignoring possible homograph markers and reflexives) based on the strict morphological criterion as adopted in MorFlex.

Reflexives. Finally, the established mapping was propagated to the reflexive verbs as well.

6. Outcome: Updated Lexicons with Interlinked Records

The two updated lexicons, MorfFlex and VALLEX, with interlinked records, represent the project's main outcome. Technically, the new attribute `-morfflex` was added to each lexical unit of VALLEX, listing all relevant MorfFlex lemmas (as representatives of the whole morphological paradigms). The total sum of 5,098 lemmas are interlinked.

Within the project, we have verified the theoretical approaches and harmonized the treatment of variants in both lexicons, adopting the clear morphologically based criteria from MorfFlex for distinguishing variants in VALLEX. As a secondary benefit, some minor inconsistencies in both lexicons were detected and corrected. In all cases, these imperfections concerned individual lexicon entries (as, e.g., missing entry was added, unrecognized variants linked, and lemmas that do not function as morphological variants were separated) while the overall design of the lexicons proved to be suitable for such endeavor.

The updated VALLEX data are publicly available in the working version.[9] The finalized version will be part of its next public release.

Since we concentrated on morphological characteristics, after analyzing the data in both dictionaries, the vast majority of lemmas were linked automatically (the few detected ambiguities were disambiguated manually as it was not worth inventing some heuristics or relying on machine learning methods for such a minor task).

Unfortunately, this does not apply to dictionaries focused on semantic information, as can be exemplified by the ambitious SemLink project for English [19, 20]. Linking such resources requires extensive manual effort to satisfy a reasonable quality of the result and large manually annotated corpora with individual predicate senses disambiguated by trained linguists. Only such training data make it possible to design elaborated (semi)automatic linking procedures allowing their users to preserve the mapping of such (necessarily constantly changing) resources. Such data are not available for VALLEX yet; however, VALLEX is being integrated into the SynSemClass project [21], which aims to serve as an inter-connecting data resource.

[9] https://quest.ms.mff.cuni.cz/vallex/

Acknowledgment

The work on the VALLEX and MorfFLex lexicons was supported by and 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).

References

[1] J. Hlaváčová, M. Mikulová, B. Štěpánková, Konzistence morfologického slovníku morfflex, Jazykovedný časopis / Journal of Linguistics 72 (2021) 855–861.

[2] J. Hajič, J. Hlaváčová, M. Mikulová, M. Straka, B. Štěpánková, MorfFlex CZ 2.0, 2020. URL: http://hdl.handle.net/11234/1-3186, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

[3] M. Mikulová, J. Hajič, J. Hana, H. Hanová, J. Hlaváčová, E. Jeřábek, B. Štěpánková, B. V. Hladká, D. Zeman, Manual for Morphological Annotation, Revision for the Prague Dependency Treebank - Consolidated 2020 release, Technical Report TR-2020-64, Institute of Formal and Applied Linguistics, Charles University, 2020.

[4] J. Hlaváčová, Golden rule of morphology and variants of wordforms, Jazykovedný časopis / Journal of Linguistics 68 (2017) 136–144.

[5] C. F. Baker, C. J. Fillmore, J. B. Lowe, The Berkeley FrameNet project, in: COLING-ACL'98: Proceedings of the Conference, Montreal, Canada, 1998, pp. 86–90. doi:https://dl.acm.org/doi/10.3115/980845.980860.

[6] M. Palmer, D. Gildea, P. Kingsbury, The Proposition Bank: An Annotated Corpus of Semantic Roles, Computational Linguistics 31 (2005) 71–106. doi:https://doi.org/10.1162/0891201053630264.

[7] K. Kipper-Schuler, VerbNet: a broad-coverage, comprehensive verb lexicon, Ph.D. thesis, Computer and Information Science Department, Universiy of Pennsylvania, Philadelphia, PA, 2005. URL: http://repository.upenn.edu/dissertations/AAI3179808/.

[8] R. Weischedel, E. H. amd Mitchell Marcus, M. Palmer, R. Belvim, S. Pradhan, L. Ramshaw, N. Xue, Ontonotes: A Large Training Corpus for Enhanced Processing, Springer New York, NY, Berlin, 2011, pp. 53–63. doi:https://doi.org/10.1007/978-1-4419-7713-7.

[9] M. Lopatková, V. Kettnerová, E. Bejček, A. Vernerová, Z. Žabokrtský, Valenční slovník

českých sloves VALLEX, Nakladatelství Karolinum, Praha, 2016.
[10] M. Lopatková, V. Kettnerová, J. Mírovský, A. Vernerová, E. Bejček, Z. Žabokrtský, VALLEX 4.5, 2022. URL: http://hdl.handle.net/11234/1-4756, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.
[11] Z. Žabokrtský, M. Lopatková, Valency Information in VALLEX 2.0: Logical Structure of the Lexicon, The Prague Bulletin of Mathematical Linguistics (2007) 41–60.
[12] J. Filipec, F. Čermák, Česká lexikologie, Academia, Praha, 1985.
[13] D. A. Cruse, Lexical Semantics, Cambridge University Press, Cambridge, 1986.
[14] P. Sgall, E. Hajičová, J. Panevová, The Meaning of the Sentence in Its Semantic and Pragmatic Aspects, Reidel, Dordrecht, 1986.
[15] J. Panevová, Valency Frames and the Meaning of the Sentence, in: P. A. Luelsdorff (Ed.), The Prague School of Structural and Functional Linguistics, John Benjamins, Amsterdam/Philadelphia, 1994, pp. 223–243.
[16] A. Rosen, Haplology of Reflexive Clitics in Czech, in: E. Kaczmarska, M. Nomachi (Eds.), Slavic and German in Contact: Studies from Areal and Contrastive Linguistics, volume 26 of *Slavic Eurasian Studies*, Slavic Research Center, Sapporo, 2014, pp. 97–116.
[17] B. Havránek, J. Bělič, M. Helcl, A. Jedlička (Eds.), Slovník spisovného jazyka českého, Academia, Praha, 1964.
[18] J. Filipec, F. Daneš, J. Machač, V. Mejstřík (Eds.), Slovník spisovné češtiny pro školu a veřejnost, Academia, Praha, 2003.
[19] M. Palmer, Semlink: Linking PropBank, VerbNet and FrameNet, in: Proceedings of the Generative Lexicon Conference, GenLex 2009, Pisa, 2009, pp. 9—-15.
[20] K. Stowe, J. Preciado, K. Conger, S. W. Brown, G. Kazeminejad, J. Gung, M. Palmer, SemLink 2.0: Chasing lexical resources, in: Proceedings of the 14th International Conference on Computational Semantics (IWCS), Association for Computational Linguistics, Groningen, The Netherlands (online), 2021, pp. 222–227. URL: https://aclanthology.org/2021.iwcs-1.21.
[21] Z. Urešová, K. Zaczynska, P. Bourgonje, E. Fučíková, G. Rehm, J. Hajič, Making a Semantic Event-type Ontology Multilingual, in: Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), European Language Resources Association, Marseille, France, 2022, pp. 1332–1334.