Obsah / Content

Články / Articles

Iterativa typu *bývat, dělávat, chodívat* v současné češtině

Ondřej Bláha
Katedra bohemistiky, Filozofická fakulta, Univerzita Palackého
ondrej.blaha@upol.cz

Habitual-iterative verbs such as *bývat, dělávat, chodívat* in contemporary Czech

ABSTRACT: The paper deals with the inventory and distribution of habitual-iterative verbs in contemporary Czech, using data acquired from corpora SYN2000 and SYN2015. These corpora are parts of *Czech National Corpus* and refer to two subsequent phases in the development of contemporary Czech, containing texts from years 1990 to 1999 and from 2010 to 2014. The data show that iterative verbs with habitual meaning recede, to a certain extent, in contemporary Czech: the corpus SYN2000 includes 475 lemmas and 40 714 tokens of iterative verbs and corpus SYN2015 414 lemmas and 49 929 tokens. This indicates that mentioned receding of iteratives in Czech refers only to their inventory (the number of lemmas), not to the occurence of these verbs, which is even slightly rising. The iterative verbs are typical for written language: they are much less common in spoken Czech. Comparing mentioned two samples, we can observe that there are some subtile changes in the stylistic distribution of the iterative verbs: these verbs prevail in specialized or administrative texts in both samples, but the younger sample (SYN2015) contains the slightly higher number of iterative verbs in journalistic texts than the older sample (SYN2000), whereas the number of iteratives is descending in ficitional texts (when compared both the samples). In addition, the iterative verbs are not so common in texts translated into Czech language from English, compared with texts that were originally written in Czech. From the point of view of grammar, the iterative verbs maintain (more or less) their grammatical particularities like the lack of the future or imperative forms and the tendency to occur in the finite verbal forms and in the 3rd person form more often than it is usual among non-iterative, normal verbs. The occurence of iterative verbs in preterite, as the younger sample shows, is rather increasing.

KEYWORDS: Czech language, verbal aspect, aktionsart, iterative verbs, current tendencies
KLÍČOVÁ SLOVA: čeština, slovesný vid, způsoby slovesného děje, iterativa, vývojové tendence

Čeština v Albrantových koňských lékařstvích

Alena M. Černá
Ústav pro jazyk český AV ČR, v. v. i., Praha
alenacerna@ujc.cas.cz

The Czech Language in Albrant's Horse Medicine

ABSTRACT: We have several manuscripts and old prints coming down to us from the 15th–18th centuries concerned with treatment of horse ailments. The Czech editions have their origins in a German treatise on horse medicine authored by the horse dealer and equerry Albrant. As years went by, they were transformed and enlarged. This article is concerned with the Czech language in the horse medicine scripts, with a particular interest in phonology and spelling (the issue of baroque vowel quantity), and also lexis (terminology, German borrowings). The material base comprises eight Czech editions of Albrant's medicine (the National Library of the Czech Republic XI C 2 and XVII E 42, the National Museum Library IV H 28, I H 29 and I F 10), a humanist print from 1527, and two undated baroque prints from the second half of the 18th century. Given this sample extending over such a large period of time, it is possible to track the Czech language evolution in its individual phases. We cannot assume, however, that the specific manuscripts and prints actually reflect the era in which the transcript of print originated. All texts involve contamination of different evolutionary phases of Czech and they provide differing amounts of the language as it appears in original texts, penetrated by linguistic elements typical of the time of transcription or printing. While analysing the language of individual texts, we focus especially on the phonological sphere, including a note on spelling and graphics, on lexis – which is rather specific (terminology) given the topical delimitation and also its origins based on the German archetype, and which is but sparsely found in other texts –, and on the overall style. Special attention is given to the vocalic quantity of the prints – it differs from the quantity assumed in the language of the humanist and baroque eras. We cannot, however, consider these deviations to be errors or print imperfections. On a thorough study of quantity markers in prints on horse medicine, it is apparent that the "different" quantity occurs in such word groups or forms that were defined also on the basis of research of other Early-Modern-Period texts (e.g. instr. sg. *tou masti*, *s soli* etc.; unification of quantity in the *hráchu*, *chléba* paradigm; adjective endings *dušny*, *psi lejno*, *nepranym* etc.; 2nd sg. imp. *znamenáj*, *vařiž*, *směžíš*, *máž*, *hás* etc.; 3rd sg. pres. ind. 2nd class *rozsedné*, *zhustné* etc.). A very significant linguistic part consists of terms denoting horse ailments. They are either of domestic (*ochvata*, *prchněly*, *záskoka*, *přístih*, *sadmo*, *šťkavka* etc.), or German origin (*šál*, *špát*, *halguf* etc.). An intertextual research has shown that the closest relationship occurs between both baroque texts, though it is not possible to determine which of the two prints may have served as a model for the other. As for the manuscripts, we do not assume that any of them would have become an immediate template for another – though the texts certainly are built around Albrant's core text, the rest of the manuscripts are further extended and varied by the interference of their later users.

KEYWORDS: history of veterinary medicine, master Albrant, historical Czech
KLÍČOVÁ SLOVA: dějiny veterinárního lékařství, mistr Albrant, historická čeština

Segmentálna a suprasegmentálna charakteristika akusticko-auditívnych komunikátov (na pozadí binárnych opozícií)

Lena Ivančová
Filozofická fakulta Univerzity Pavla Jozefa Šafárika v Košiciach
lena.ivancova@upjs.sk

Segmental and Suprasegmental Characteristic of Acoustic-Auditory Communication Units (On the Background of Binary Oppositions)

ABSTRACT: The paper introduces a systematic way of segmental and suprasegmental characteristic of acoustic-auditory communication units on the background of seven binary oppositions. The basic, neutral element of observed acoustic characteristic and its marked opposite are defined within each opposition. The described system of the acoustic characteristics of phonemes and intonemes is based on the acoustic parameters of the Slovak language, however, it can also be applied as a model starting point in describing acoustic-auditory communication units in other languages.

KEYWORDS: spoken discourse, binary oppositions, intonation, phonetics

KĽÚČOVÉ SLOVÁ: hovorený prejav, binárne opozície, intonácia, fonetika

Dialogické rysy dopravního rozhlasového zpravodajství (na příkladu slovesných způsobů)

Lucie Jílková
Ústav pro jazyk český AV ČR, v. v. i., Praha
jilkova@ujc.cas.cz

Dialogical features in traffic radio news (on the example of verbal moods)

ABSTRACT: The article deals with the analysis of radio news on traffic. Traffic news is generally described as a specific kind of a dialogue, namely the dialogue between a reporter and phoning drivers, acted out as a role-play. This description is based on recordings from the public radio station Radiožurnál and the private radio station Impuls (10 recordings from each radio). Further attention is paid to the verbs used in traffic news, especially to those in the form of 2nd person plural in the indicative (*zdržíte se* / *you are late*) and 2nd person plural in the imperative (*čekejte*/*wait*) that is to those verb forms which are used by the reporter to address radio listeners (drivers) directly. The analysis showed that the distribution of verbal means (the indicative and the imperative) is practically identical in both radio stations.

KLÍČOVÁ SLOVA: dopravní rozhlasové zpravodajství, dialogické rysy, slovesa

KEYWORDS: traffic radio programme, features of dialogue, verbs

Původci sdělení v souvětích s větami obsahovými a jejich vliv na užívání absolutních a relativních časů

Marta Koutová
Ústav pro jazyk český AV ČR, v. v. i., Praha
koutova@ujc.cas.cz

The originators of the message in the sentences with content clauses and their impact on the use of absolute and relative tenses

ABSTRACT: Based on the analysis of the Czech National Corpus, we have found that, in addition to the assumed use of tenses (relative in content clauses, or rather, absolute in adjunct clauses), there are also variations in their use; in content clauses, absolute tenses are used in certain cases, and, on the other hand, the relative tense may be used in some of the adjunct clauses. Using the corpus evidence in this article, we show that the use of verb tenses in subordinate content clauses and all adjunct clauses explicitly or implicitly dependent on them is ruled by the speaker's point of view: whether they look at the action of the content clause from their perspective, or whether they adopt the perspective of the primary agent. If a tense is used from the speaker's perspective, it is an absolute tense, in case of the primary agent's perspective, it is a relative tense.

KLÍČOVÁ SLOVA: vedlejší věta obsahová, vedlejší věta doplňovací, relativní čas, absolutní čas, prvotní konatel, mluvčí

KEYWORDS: subordinate content clause, subordinate adjunct clause, relative tense, absolute tense, primary agent, speaker

Překlad prostředků mluvenosti v beletrii

Lenka Mundevová: Překlad prostředků mluvenosti v beletrii. Stoletá historie překladu Maupassantovy povídky L'Ivrogne. Praha: FF UK, 2018, 211 s.

František Štícha
Ústav pro jazyk český AV ČR, v. v. i., Praha
sticha@ujc.cas.cz

Tato pozoruhodná monografie, přinášející množství cenných poznatků nejen z teorie a praxe překladu obecně a z vývoje překládání jedné povídky G. De Maupassanta, ale i z vývoje pronikání obecné češtiny do prozaických textů v průběhu 20. století, má pět hlavních kapitol: 1. Stratifikace národních jazyků v kontextu překladu; 2. Metodologie: vymezení kritérií pro hodnocení překladů; 3. Normy v dějinách českého překladu od doby lumírovců do devadesátých let 20. století; 4. Kontext české literární tvorby; 5. Translatologická analýza: literární stylizace Maupassantova dialogu v originále a v překladech.

V první kapitole autorka jednoduše a výstižně připomíná, že dělení jazyka na spisovný jazyk, interdialekty a dialekty je „základem všech představ o stratifikaci češtiny". V dobrém přehledu „nejzákladnějších znaků obecné češtiny" nalézám jednu, patrně nepozornostní, chybu a jeden jev diskutabilní: Chybou je, že do „roviny slovotvorné" je zařazeno i užívání slov přejatých (*furt*, *flaška*), diskutabilní je tvrzení, že „Ke znakům mluvenosti […] patří i užívání **nefunkčních ukazovacích** a neurčitých **zájmen**." Pokud má autorka na mysli běžné případy jako ***To*** *maso je v mrazáku* nebo *Kdy půjdeme do* ***toho*** *kina?*, pak bych ji rád upozornil, že v těchto a mnoha jim podobných případech je užívání odkazovacího zájmena *ten* velmi funkční. Existuje o tom dosti literatury.

Přenesu se teď až do kapitoly třetí, neboť k předcházejícím stránkám, bohatě informativním, nemám co dodat. Na straně 48 autorka kriticky, avšak zdrženlivě cituje slova O. Fischera, že „z některé básně průměrné teprve překladem stane se tvůrčí čin". To je sice, dodávám, jistě možné, avšak čtenář překladu pak čte více překladatele než autora samého. A dále na tutéž notu zmiňuje autorka slova B. Mathesia o tom, že dobrý překladatel může a má autora znásilnit, zkrátit, prodloužit, doplnit či překomponovat. Tento postoj, podotýkám k tomu, je dnes neudržitelný a téměř neuvěřitelný, neboť většina dnešních překladatelů, teoretiků překladu, lingvistů i samotných čtenářů má jistě názor zcela opačný.[1]

Jádrem knihy je kapitola pátá, v níž se autorka podrobně věnuje fonologickým, morfologickým, syntaktickým, lexikálním a dialektálním prostředkům Maupassantova originálu a jejich překladům do češtiny; jde o překlady Pavla Projsy z r. 1902, Františka Sekaniny z r. 1920, Viléma Opatrného z r. 1950, Luďka Kárla z r. 1961 a Dany Melanové z r. 1997.

O jazyce Maupassantovy povídky autorka říká, že „z přímé řeči, jež obsahuje nářeční prvky, lze vyčíst místní a sociální zařazení postav" (s. 81), avšak její syntax nevykazuje větší odchylky od běžné komunikace a dialekt užívá Maupassant spíše jen náznakově.

Do kapitoly o nejstarším překladu Projsově autorka účelně a zajímavě vložila exkurs o dobové recepci Maupassantova díla u nás. Jde o dobové články publikované v časopisech *Ruch*, *Lumír*, *Rozhledy literární*, *Národní listy*, *Světozor*, *Obzor*, *Našinec*, *Moderní Revue* a *Plzeňské listy*. Z Národních listů z r. 1888 např. autorka cituje anonymního recenzenta, jenž chvále Maupassantovu výrazovou střídmost píše: „Snažme se býti výbornými ovladateli slohu a nikoli sběrateli neobyčejných slovíček." Tento výrok je u národa, jenž si svou uměleckou prózu teprve budoval, sice pochopitelný, avšak z dnešního hlediska je naivní a nepřijatelný. Vždyť co „neobyčejných slovíček" bychom našli nejen u Vladislava Vančury nebo Karla Čapka. Jiný tehdejší kritik zase píše: „Bez balastu a zbytečných slov […] líčil Maupassant prostý život, nahou pravdu […]." To jsou ovšem typické výroky kritiků nejen 19. století, které se snadno píší a hezky čtou, jejichž pravdivost je ale zahalena čirou tmou. Kdo dokáže, co jsou to zbytečná slova a proč jsou zbytečná? Kdo dokáže rozlišit slovní balast od slovního skvostu? Podobně psal o Maupassantovi, jak se dozvídáme díky autorce, i Jaroslav Vrchlický: „On nikdy se nezabere v absurdnosti neologismu […]." Jak absurdní výrok básníkův!

Pokud jde o jazyk Projsova překladu, autorka upozorňuje např. na výrok Šaldův o tom, že „Projsa si libuje […] v barokně uspořádaných větách a ponechává Maupassantovi vlastně jen fabuli." A na výrok z *Lexikonu české literatury*, že Projsa nedbal „na slohovou specifičnost originálu".

Na s. 103 autorka definuje výrazy, které pokládá za hovorové. Jsou to nejen ty, které slovníky (PSJČ, SSJČ) hodnotí jako obecněčeské, hovorové, expresivní, hanlivé, zhrubělé apod., ale i takové, u nichž slovníky žádnou značku stylové charakteristiky neuvádějí, ale jejichž hovorovost „jasně vyplývá z kontextu".

Přestože autorka při hodnocení jazyka překladu postupuje velmi precizně krok za krokem, některé detaily, o nichž se nezmiňuje, by byly stály za zmínku. Např. výraz *podíváš se* v Projsově větě „Řekni mi, co to bylo Mélino, nebo tě sbiju, **podíváš se**!" je z dnešního hlediska nesrozumitelný. V originále je přitom výraz, který dosti věrně odpovídá českému *varuju tě* – tohoto výrazu také užila Dana Melanová. Po zhodnocení všech dílčích prostředků fonologických, morfologických, syntaktických a lexikálních autorka dospívá k závěru, že Projsův překlad působí téměř spisovně, že „prvky mlu-

[1] Jde tu obecně o otázky věrnosti překladu. K nim viz F. Štícha: *O věrnosti překladu*. Praha: Academia, 2019.

veného jazyka jsou v Projsově překladu méně frekventované než v originále“ a že „překlad vykazuje nižší míru expresivity než předloha“ (s. 111). Těžištěm mluvenosti v překladu jsou přitom podle autorky prvky syntaktické a zejména lexikální, **mluvené lexikum je v překladu zastoupeno více než v originále**. To je, dodávám k autorčině hodnocení, z hlediska dnešní češtiny velký paradox, neboť zatímco v ostatních jazycích se dnes mluvenost promítá právě do lexika a syntaxe – neboť do morfologie se téměř nemá kam promítat – dnešní mluvená čeština se jen hemží morfologickými tvary, jež jsou psané spisovné češtině zcela cizí. Příčina toho je obecně známá a autorka to jistě dobře ví, ale ve svých hodnoceních na to zřejmě někdy pozapomněla.

Autorka Projsovi neprávem vytýká, že „příznakovost originálu se zde nijak nepromítla do roviny morfologické a fonologické. Jde o větu *To on mě zdržel u toho halamy břicháče Paumella a jako každý večer… abych se nedostal domů. Takhle to vypadá! Něco provádí, odranec, lotr!* Ale jak se tu mohla mluvenost a nespisovnost na samém počátku 20. století promítnout do morfologie? Tehdejší překladatel stěží mohl napsat *von* místo *on*, *každej* místo *každý* a *domu* místo *domů*. A kde vzít pro tuto větu jiné hovorové a nespisovné morfologické prostředky? Autorka si to snad dodatečně uvědomila, neboť na téže stránce o něco níže píše, že překladatel se patrně snažil „částečně kompenzovat nedostatek prostředků mluveného jazyka na fonologické a morfologické rovině.“

Stejně jako u Projsy řadí autorka i u F. Sekaniny a v dalších překladech mezi prostředky mluvené češtiny i „ukazovací zájmena a další zástupné výrazy s obecnou platností“, např. ***To*** *on mě teda zdržel, u* ***toho*** *ničemy Paumella* a říká o nich, že „jejich užití vede ke zjednodušení výpovědi“. Tomuto hodnocení nerozumím a nikde v lingvistické literatuře jsem se s ním nesetkal.

„Stejně jako text Projsův,“ píše autorka, vykazuje i text Sekaninův „nižší míru mluvenosti než originál“. „Těžištěm mluvenosti Sekaninova překladu je lexikum. Lexikální prostředky jsou zde rozmanitější než u Projsy,“ tvrdí autorka (s. 127). A konečně: „Sekanina lne k významové věrnosti více než Projsa.“ Také Sekaninovi však autorka vytýká, že Maupassantův jazyk nivelizuje, když nevyužívá „koncovek typických pro obecnou češtinu“. To však, namítám, v tehdejší době nebylo dost dobře možné.

O zacházení s nespisovnými prostředky v překladu Viléma Opatrného autorka shrnujícím způsobem říká, že se v něm „mísí spisovné až knižní prostředky s nespisovnými“ (s. 143). Z jednotlivostí mě tu nejvíce zaujal výraz „suvák“ (z francouzského *sou*), o němž jsem předpokládal (neboť jde o unikátní slovotvorný solitér!), že SSJČ ho nebude registrovat, ale mýlil jsem se. Slovník tento výraz registruje a uvádí jeden doklad jeho užití z díla Marie Majerové (žila delší dobu v Paříži). Autorka recenzované knihy o tomto výrazu nic neříká, což je škoda, neboť čtenáři neznalému francouzštiny a nemajícímu bližší vztah k francouzské literatuře a k francouzským reáliím, toto slovo nejspíš nic neřekne.[2]

[2] Předpokládám, že tuto knížku nebudou číst jen romanisté nebo jen lingvisté či překladatelé a teoretici překladu znalí francouzštiny.

Shrnujíc výsledky překladatelské metody Luďka Kárla, autorka říká, že „Ten se co do počtu užitých prostředků mluveného jazyka (99) nejvíce přiblížil originálu (110)“, a patrně správně k tomu dodává, že svou roli tu sehrála „dobová estetická norma“ a jí odpovídající „stoupající odvaha překladatelů“ užívat prostředky mluveného jazyka.

V překladu Dany Melanové si autorka všimla toho, že překladatelka důsledně podržuje spisovné podoby s dlouhým vokálem ve tvarech jako *platím, prosím, vím* i v „neformálně laděných“ dialozích. Podotýká k tomu, že překladatelka „nechtěla ‚zahltit‘ povídku prostředky nespisovnosti, čímž by z ní setřela rys staršího textu […]“ (s. 170).

Lenka Mundevová se ve své užitečné knížce věnuje také některým nepřesnostem, nevěrnostem či vyloženým chybám v překladech. Více ale už čtenáři této recenze neprozradím. Nechť si v této knížce sám počte a vybere si z ní to, co zaujme jeho samého.

VerbaLex – Comprehensive Dictionary of Czech Verb Valencies

Dana Hlaváčková, Aleš Horák, Karel Pala
NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic
{hlavack, hales, pala}@fi.muni.cz

In this paper, we deal with the Czech valency dictionary named VerbaLex. It is a lexical database containing 10 449 Czech verb lemmata with their valency frames providing morphosyntactic and semantic information about the verb arguments. The verbs are organized as synonymical sets and linked to the Princeton WordNet. In this respect VerbaLex differs from the other Czech valency dictionary named Vallex and developed in UFAL, Prague. The number of the valency frames in VerbaLex is about 19 500. They include information about various properties of the Czech verbs such as surface cases, reflexivity, aspect, or references to the English translational equivalents via ILI.

1. Introduction

In natural languages verbs represent the part of speech, which typically serves for organizing elements of the sentence thanks to their argument predicate structure. Thus it is desirable and useful to create lists of verbs capturing their properties, especially their valencies indicating how verbs are binding other parts of speech, most frequently nouns. Usually, we speak about valency dictionaries that have been becoming more popular recently. For English the VerbNet lexical database should be mentioned (https://verbs.colorado.edu/~mpalmer/projects/verbnet.html).

For Czech there are three valency dictionaries, the first one is Vallex prepared in UFAL MFF UK (http://ufal.mff.cuni.cz/vallex), the second one, named *VerbaLex*, was developed in NLP Centre FI MU. The third one is *Slovesa pro praxi* (Svozilová et al. 1997).

The Czech language can be considered by researchers as a sort of fortress they are trying to attack from different angles. Thus *VerbaLex* offers a distinct view of the Czech verb valencies than Vallex – the main disparity lies in the notion of semantic roles characterizing the meaning of the verb arguments (in Vallex named functors – their number is 32). Semantic roles in *VerbaLex* are inspired by the Top Ontology developed in the framework of the EuroWordNet and Base Concepts in it, thus they display two levels – main roles (38 items) and the selected WordNet nodes (literals) that can be characterized as selectional restrictions. Their number is approximately 811 items and their list is open (see below Sect. 3.1).

VerbaLex can serve for expert public, linguists, students, translators, researchers in the NLP area and anyone who may be interested in obtaining a deeper knowledge about syntax and semantics of Czech. Therefore, it can also be used as a data resource in various computer applications such as syntactic analysis, information search, summarization, as well as machine translation.

2. The VerbaLex Valency Dictionary

VerbaLex is an electronic lexical database comprising verb valency frames developed in the NLP Centre, Faculty of Informatics, Masaryk University in Brno (FI MU) during 2006–2013. It is a result of the work, belonging partly to the area of linguistics and partly to the field of Natural Language Processing (NLP). During the development of *VerbaLex*, we have been using various corpus and electronic resources, which made it possible to observe the behaviour of the verbs in their natural contexts. The main part of the database has been compiled by the annotators who relied on their linguistic competence, followed given instructions and using the accessible software tools they created what is called the *basic* and *complex valency frames*.

The verb valency is understood here as a semantically given ability of the verb allowing it to combine with other words – the verbs are described from this point of view together with their complements both on their left and right side. Thus valency frames contain two kinds of information: the *morphosyntactic* and *semantic* one. Our effort was to capture as many Czech verbs as possible: presently the *VerbaLex* comprises 10 449 verbs. For compiling *VerbaLex* we have used some existing resources, in the first place the *Valenční slovník českých sloves* (*Valency Dictionary of Czech Verbs*) also known with the working name *BRIEF* (Pala, Ševeček 1997) and containing approximately 15 000 Czech verbs with their surface valency frames.

Our motivation has been an effort for a deeper understanding of the semantics of Czech verbs and their arguments and creation of the new electronic data resource. In comparison with the traditional approaches, for instance (Svozilová et al. 1997), we have used methods and techniques, particularly semantic networks and ontologies, which do not appear in the existing Czech dictionaries at all. The obtained results can be then exploited in the field of NLP, since *VerbaLex* captures relevant semantic relations.

2.1 VerbaLex Composition

The database displays some basic features, in which it differs from similar dictionaries. The form of the *complex valency frames* allow us to capture the relevant information about a verb and its complements. The valency frames are assigned to individual verb senses (grouped in synonymical sets or *synsets*, see Section 2.4) and not only to individual lemmata (many synonyms share the same valency). To label the meanings of the verb complements the concept of the two-level semantic roles has been developed.

The *basic valency frames* (see Figure 1), which represent the core of *VerbaLex*, constitute the notation of the verb valency on the morphosyntactic and semantic level. The center of the frame is a marked verb position, its valency complements on the morphological level are represented by the pronominal expressions together with the respective case numbers. The notation follows the canonical word-order: the complement on the left side – verb – the complements on the right side. On the semantic level the verb arguments are labeled by the two-level semantic roles, which specify the semantic environment of the verb as precisely as possible. The frame contains additional information about obligatoriness and optionality of the valency complements. The basic valency frame is always related to a *subsynset*, which is a subset of the defined synonymical set.

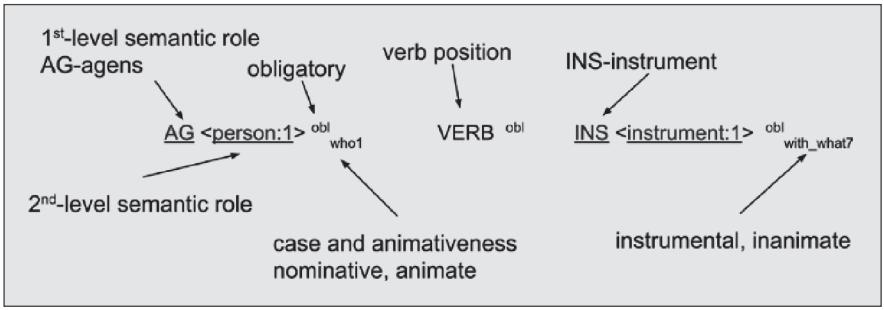

Figure 1: Basic valency frame for: Žák píše perem. (A pupil writes with pen.)

The basic frame is a part of the complex valency frame (see Figure 2), which is always related to the one synonymical set only. Apart from the frame there is additional information which includes the verb sense, aspect (see Section 2.3) and verb semantic class (see Section 3.2). For each verb its ability to form passive voice is marked, thus it is possible to obtain lists of the transitive and intransitive verbs from the database. The important feature of the Czech verbs is their obligatory or optional reflexivity – we provide information about three basic types of reflexivity: proper reflexives (reflexiva tantum), i.e. verbs with obligatory reflexive particle *se*, e.g. *bát se* (*to be afraid*). Further, object reflexivity is marked when the pronoun *se* or *si* replaces object of the action (*mýt se*, *čistit si* (*wash yourself*, *clean yourself*)) as well as reciprocity, mutual activity of the two subjects (*znát se*, *milovat se* (*to know each other*, *to love each other*)). Here the verb lemma is not given with the reflexive pronoun. For the rest of the verb lemmata, we mark the fact that they have or do not have the reflexive form in the respective sense and characterize it as another (not specified) type of the reflexivity. The next relevant information is related to the behaviour of the verb in a particular context: we mark their primary (basic) usage in contrast to the figurative (metaphorical) meaning. In some cases displaying a higher frequency in corpora we also indicate the idiomatic usage.

cestovat:1 (impf), **jezdit:4** (impf), **putovat:4** (impf) (travel:2, journey:1)
definition: *podnikat cesty a výpravy (undertake a journey or trip)*
class: *run-51.3.2*
passive: *no*
-frame:AG(who1;<person:1>;obl) VERB LOC(to_what2;<positon:1>;opt)
INS(by_what7;<vehicle:1>;opt)
-example: *cestuje trajektem do Dánska (impf) (he travels by ferry to Denmark)*
-use: *prim*
-reflexivity: *no*

Figure 2: Complex valency frame

2.2 Creation of the Verb List

The choice of the verb lemmata contained in the database *VerbaLex* has been based mainly on *Slovník spisovné češtiny* (Dictionary of Literary Czech [SSČ 2005]) and *Slovník spisovného jazyka českého* (Dictionary of Literary Czech Language [SSJČ 1989]). Moreover, the lemma selection stylistic features and frequencies of the particular verbs in the corpora *SYN2000* [6] and *Czes2* (*NLP Centre FI MU, 465 mil. tokens)* have been taken into consideration. As a basic resource, the *Valenční slovník českých sloves*, or *BRIEF* (Pala, Ševeček 1997), has served describing surface right side valency complements (without information about their meaning) for 15 079 Czech verbs. In *VerbaLex*, we have stored only verbs belonging to the literary Czech, where some of them can eventually have the emotional colouring. *VerbaLex* does not comprise verbs from colloquial Czech and dialects as well. We also have left aside verbs that are strongly bookish, archaic or rarely used. However, we have taken into consideration the cases when a verb is marked in a dictionary as bookish or rarely used (e.g. *pravit* (*say*, bookish form)), but they show high frequency in the corpora (verbs like *pravit* (*to say*) – shows 29 397 occurrences in the corpus *Czes2*).

The verb lemma is always one-word, in the case of the proper reflexives (reflexiva tantum) with the reflexive particles *se*, *si*. The database does not contain negated forms of the verb lemmata, we work with the assumption that the valency frames of the negated verbs remain unchanged (except for cases with the negated genitive) and the negated forms of lemmata it is possible to derive automatically. In *VerbaLex*, there is a formal way how to handle to what we call variant lemmata. In such cases the verb forms differ only in the vowel alternation, otherwise all their characteristics remain the same (e.g. *muset*/*musit* (*must*), *bydlit*/*bydlet* (*live*), *červavět*/*červivět* (*become wormy*)).

2.3 Synonymy and Verb Senses

As we have indicated above, verbs in *VerbaLex* are organized in synonymical sets (*synsets*). In synsets, each verb lemma (and its variants) are marked with the ordinal

number denoting their sense.[1] This marking directly corresponds to the numbering in the Czech WordNet, see Section 3. The appropriate synonyms have been chosen and verified in *Slovník českých synonym* (Dictionary of Czech Synonyms [SČS 1996]). Each synset is accompanied with a short description of its meaning. The descriptions (with necessary modifications) are formulated on the basis of the lexicographic definitions in *SSČ* and *SSJČ*. In the specification of particular verb senses, we often cannot always use the *SSČ* and *SSJČ* directly. Their approach differs in many cases and they contain different number of verb senses. These dictionaries are also not sufficiently up-to-date source of the current language. In case of a verb sense, which was not found in the dictionaries, the verb occurrences in new contexts were verified on corpus data (*SYN2000 and Czes2*). If the number of the verb sense occurrences reached adequate frequency, the sense was added to *VerbaLex* with a new sense number. Obsolete and rare cases from the dictionaries are not used in *VerbaLex* at all. In accordance with the *WordNet* structure,[2] the verb sense determination is often more fine-grained than in usual Czech dictionaries.

2.4 Capturing Verb Aspect

VerbaLex contains a formal notation of the verb aspect as related to the respective verb sense. The aspect identification is primarily based on the information in the dictionaries *SSČ* a *SSJČ*. If a given verb sense is valid for both its aspect forms, the verb in the first position is marked as perfective (*pf*) together with the number of the sense followed in the brackets by its imperfective form (*impf*), which automatically takes over the sense number from the perfective and it is not necessary to indicate it again. The verbs with two aspects are denoted as biaspectual (*biasp*). Iterative verb forms are not stored in the database, their forms can be automatically added from the existing morphological database to the *VerbaLex* at any time.

3. VerbaLex and WordNet

One of the main features of *VerbaLex* is a close relation to the *WordNet* semantic network (*Princeton WordNet*, *PWN* [Fellbaum 1998]). During the *BalkaNet* EU project in 2002[3], the Czech WordNet (*CzWn*) structure was supplemented with the basic valency frames including semantic roles as developed in the EuroWordNet project (see 3.1). According to the *PWN* structure, the frames were linked to whole synsets instead of individual verb lemmata. For the same reason, the frames were divided according to particular verb senses.

The verb synonymy is understood here in a broader sense than usual. The synset participants are often *near synonyms*, which cannot be freely interchanged in the

[1] The tuple "lemma:sense" is often called a *literal*.

[2] see Section 3

[3] http://www.dblab.upatras.gr/balkanet/

same contexts. In the late 90s, the *PWN* approach was applied within the EU projects *EuroWordNet-1* and *2* (*EWN*), in which new national *WordNets* were created for Dutch, Italian, Spanish, French, German, Czech and Estonian. The synsets in the national *WordNets* were interlinked by means of *Interlingual Index* (*ILI*) describing the translational equivalents. In each language, for which a *WordNet* was created, we can find at least 15 000 synsets with equivalents in *PWN*.

In *VerbaLex*, all verb senses are directly linked to their English equivalents in *PWN*. The newly added synsets were linked to the *PWN* English synsets using the WordNet Assistant tool (Němčík et al. 2008). Appropriate equivalents could be found for 85 % out of 3 686 new synsets. In 15 % the direct lexicalized equivalent could not be found – for perfective, reflexive or prefixed verbs, or verbs with expressive or metaphoric meaning.

3.1 Semantic Roles

Within the *EWN* projects, the core of the shared interlingual lexicon was defined by means of the *Top Ontology* and a larger set of *Base Concepts*.[4] The top ontology was also inspiring for the *VerbaLex* system of two-level semantic roles. Above all, we have selected the concepts covering large classes of lexical meanings. The classes correspond to the top hypernyms in the *PWN* hierarchy. We have chosen the hypernyms that best reflect the relevant meanings of the semantic roles and that are branching to expected hyponyms. *VerbaLex* 1st-level semantic roles use literals with sense number 1 or 2, i.e. basic meanings, which belong to the set of Base Concepts. The whole set of 1st-level roles is currently formed by 38 main semantic roles, which describe very general meanings taken from the Top Ontology. Each role covers one well recognized and specified meaning area, e.g. *ARTifact*, *ACTivity*, *INStrument*, *COMmunication*, *EVENt*, *LOCation*, or *TIME*.

Instrument – in *VerbaLex* a semantic role:
1st level – INS
2nd level, PWN hypernym – **instrument:1**
Two-level semantic role – INS(instrument:1)
Hyponymic lexical units as specifiers:
INS<computer:1>, *INS<weapon:1>*, *INS<sports equipment:1>*, *INS<cutlery:2>*, *INS<musical instrument:1>*, …
Hyponymic subclass of particular examples:
INS<weapon:1> = gun:1, sword:1, knife:2, bow:4, spear:1, …

Figure 3: Example of a two-level semantic role

[4] At the beginning the set included about 1000 base concepts, which was later extended to 8000 concepts.

The 2nd-level roles use direct hyponyms from *PWN* serving as a specification of the "most expected" main meaning of the verb argument. The hyponyms of such literals can then serve as instances of the appropriate class. An example can be two-level roles denoting all instruments, as shown in Figure 3. The 2nd-level roles can also be understood as subcategorization features, or selectional restrictions. They form an open system of labels, which can be continuously extended with regard to current applications.[5] The motivation for such approach lies in the aspiration to obtain a detailed description of the particular verb senses.

3.2 Semantic Classes of Verbs

The *VerbaLex* database contains not only meanings of the verb arguments, but also the meaning of the verb itself, which is one of the principal factors of its valency frames at both syntactic and semantic levels. The verb meaning is, besides the human readable definition, captured by detailed classification using verb semantic classes. Experimentally, we have chosen the classification system of English verbs by B. Levin (Levin 1993), which builds upon the syntactic and semantic features of English verbs. The system divides verbs according to alternations of their participants. Within the *VerbNet* project of M. Palmer 48 basic semantic classes of Levin were extended to 83 classes (numbered 9–91. [Palmer et al. 1998]). Ambiguous verbs, originally instantiated in multiple classes, were detached to individual classes with their own meaning.

In *VerbaLex*, we have adapted the original set of semantic classes from *VerbNet* (numbered from 9 according to Levin and Palmer) and divided some of them to meaning subclasses resulting in 109 current verb class collections. The classes are originally based on the description of changes in the argument structure of English verbs, but after the appropriate adaptation they can serve to the purpose very well also for Czech.

4. Conclusions

In this paper, we have presented the comprehensive valency dictionary of Czech Complex valency frames named *VerbaLex*. With its 10 449 verb lemmata, *VerbaLex* represents the largest Czech verb valency dictionary aiming at machine processing of verb frames in Czech sentence analysis.

VerbaLex Complex Valency Frames offer the both morphosyntactic and semantic information about Czech verbs:

- definition of the verb meanings for each synset;
- verb ability or inability to create passive form;
- number of meanings for homonymous verbs;
- semantic class a verb belongs to;

[5] *VerbaLex* currently contains 811 2nd level semantic roles.

- verb aspect (perfective, imperfective, biaspectual);
- types of verb use (primary, figurative, idiomatic);
- types of reflexivity for reflexive verbs.

VerbaLex exists in electronic form, which is accessible online after registration for academic and non-commercial purposes (see https://nlp.fi.muni.cz/verbalex/htmlDEMO/) and also http://nlp.fi.muni.cz/declaration/).

REFERENCES

FELLBAUM, C. (ed.) (1998): *WordNet: An Electronic Lexical Database. Language, Speech and Communication*. Cambridge: MIT Press.

FILIPEC, J. – DANEŠ, F. – MEJSTŘÍK, V. (eds.) (2000): *Slovník spisovné češtiny pro školu a veřejnost*. Prague: Academia.

HAVRÁNEK, B. et al. (1989): *Slovník spisovného jazyka českého*. Prague: Academia.

Institute of Czech National Corpus, F. C. (2000): *Czech national corpus* {syn2000 http://ucnk.ff.cuni.cz.

LEVIN, B. (1993): *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago: University of Chicago Press.

LOPATKOVÁ, M. – ŽABOKRTSKÝ, Z. (2008): *Vallex*, http://ucnk.ff.cuni.cz, 6460 lexical units.

NĚMČÍK, V. – PALA, K. – HLAVÁČKOVÁ, D. (2008): Semi-automatic linking of new Czech synsets using Princeton Wordnet. In: *Proceedings of the Intelligent Information Systems XVI Conference (IIS'08)*. Warszawa: Academic Publishing House EXIT, 369–374.

PALA, K. – ŠEVEČEK, P. (1997): Valence českých sloves. In: *Sborník prací Filosofické fakulty Masarykovy university*, A45, Brno, 41–54.

PALA, K. – VŠIANSKÝ, J. (1996): *Slovník českých synonym*. Prague: Lidové noviny.

PALMER, M. – ROSENZWEIG, J. – DANG, H. T. – KIPPER, K. (1998): Investigating regular sense extensions based on intersective Levin classes. In: *Proceedings of the 17th international conference on Computational linguistics*. Association for Computational Linguistics, 293–299.

SVOZILOVÁ, N. – PROUZOVÁ, H. – JIRSOVÁ, A. (1997): *Slovesa pro praxi: Valenční slovník nejčastějších českých sloves*. Prague: Academia.

ŠTÍCHA, F. (2018): Valenční slovník českých sloves (Valency Dictionary of Czech Verbs). *KGA*, 18/2018, 75–81.

ACKNOWLEDGEMENTS

This work has been partly supported by the Ministry of Education of CR within the LINDAT-Clarin infrastructure LM2015071 and OP VVV project CZ.02.1.01/0.0/0.0/16_013/0001781.