Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 3136–3144
Marseille, 11–16 May 2020

Towards a Semi-Automatic Detection of Reflexive and Reciprocal Constructions and Their Representation in a Valency Lexicon

Václava Kettnerová, Markéta Lopatková, Anna Vernerová, Petra Barančíková
Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics
Malostranské náměstí 25, 118 00 Prague, Czech Republic
{kettnerova, lopatkova, vernerova, barancikova}@ufal.mff.cuni.cz

Abstract

Valency lexicons usually describe valency behavior of verbs in non-reflexive and non-reciprocal constructions. However, reflexive and reciprocal constructions are common morphosyntactic forms of verbs. Both of these constructions are characterized by regular changes in morphosyntactic properties of verbs, thus they can be described by grammatical rules. On the other hand, the possibility to create reflexive and/or reciprocal constructions cannot be trivially derived from the morphosyntactic structure of verbs as it is conditioned by their semantic properties as well. A large-coverage valency lexicon allowing for rule based generation of all well formed verb constructions should thus integrate the information on reflexivity and reciprocity. In this paper, we propose a semi-automatic procedure, based on grammatical constraints on reflexivity and reciprocity, detecting those verbs that form reflexive and reciprocal constructions in corpus data. However, exploitation of corpus data for this purpose is complicated due to the diverse functions of reflexive markers crossing the domain of reflexivity and reciprocity. The list of verbs identified by the previous procedure is thus further used in an automatic experiment, applying word embeddings for detecting semantically similar verbs. These candidate verbs have been manually verified and annotation of their reflexive and reciprocal constructions has been integrated into the valency lexicon of Czech verbs *VALLEX*.

Keywords: valency lexicon, reflexivity, reciprocity

1. Introduction

Reflexivity and reciprocity attract much attention in current theoretical linguistics, particularly from a typological perspective, see e.g. (Nedjalkov, 2007; König and Gast, 2008; Evans et al., 2007; Frajzyngier and Walker, 2000a; Frajzyngier and Walker, 2000b). These language phenomena are typically encoded by *reflexive markers* (henceforth *RM*s). However, cross-linguistic studies evidence that *RM*s typically serve various functions in a language, crossing boundaries of reflexivity and reciprocity, see esp. (Geniušienė, 1987) and (Kemmer, 1993).

In Czech, which is the focus of this work, *RM*s are involved in various constructions. In addition to reflexive (1) and reciprocal constructions (2), *RM*s encode impersonal passive constructions (3), middle constructions (4), anticausative constructions (5) or constructions with inherently reflexive verbs (6).[1]

(1) *Petr myslel jen na sebe.*
Peter thought only of RM
'Peter thought only of himself.'

(2) *Rodiče si (vzájemně) pomáhali.*
parents RM (mutually) helped
'Parents helped each other.'

(3) *Z České republiky se elektřina vyvážˊí.*
from Czech republic RM electricity exports
'Electricity is exported from the Czech republic.'

(4) *Špatně se mi spalo.*
badly RM me slept
'I did not sleep well.'

(5) *Okno se rozbilo.*
window RM broke
'The window broke.'

(6) *Petr se smál od srdce.*
Peter RM laughed heartily
'Peter laughed heartily.'

The high ambiguity of *RM*s complicates their description, often leading different theories to contradictory conclusions. Moreover, it produces many inconsistencies in their annotation, see (Marković and Zeman, 2018), making their cross-linguistic comparison and using the data in NLP tasks difficult. Lexical resources allowing for the disambiguation of various functions of *RM*s are thus highly beneficial for both theoretical linguistic research and NLP applications. Moreover, high frequency of *RM*s stresses the urgency of this task.

In this paper, we aim at filling this gap by developing a method making possible to disambiguate reflexive and reciprocal constructions in Czech from the other types of constructions with *RM*s in corpus data. Further, a list of verbs creating these constructions is compiled, forming a lexical stock in the annotation of reflexivity and reciprocity in a valency lexicon.

Reflexivity and *reciprocity* represent typical phenomena at the semantics-syntax interface: they are associated with systemic changes in the surface syntactic behavior of verbs; however, at the same time they are semantically conditioned. Thus their representation must rely both on grammatical rules underlying the surface syntactic structure of these constructions and on the information on the possibility of individual verbs to create these constructions.

Due to rather general semantic qualities conditioning the possibility of verbs to create reflexive and/or reciprocal

[1] Similar functions are exhibited by *RM*s in other Slavic and Romance languages as well (Medová, 2009).

constructions, it is not straightforward to determine the set of verbs creating these constructions. For this purpose, we propose a two-step semi-automatic procedure detecting the verbs that are likely to create the constructions under scrutiny: the first step detects candidate verbs in an large corpus relying on lexical and grammatical constraints blocking ambiguity of *RM*s and the second step consists in identification of other semantically similar verbs, using a continuous vector space representation. As a result, a list of highly plausible verbs for further manual annotation of reflexivity and reciprocity is obtained. Finally, for the obtained set of verbs, an economical and linguistically adequate representation of reflexivity and reciprocity is proposed. This representation is integrated in the valency lexicon of Czech verbs *VALLEX*.

2. Reflexivity and Reciprocity in Lexical Resources

Despite being semantically conditioned (and thus not applicable to all lexical units), reflexivity and reciprocity are mostly treated as productive processes the description of which entirely relies on grammar alone. As a consequence, explicit representation of reflexivity and reciprocity is still missing in most contemporary lexical resources. However, being associated with systemic changes in syntactic patterns of verbs, the information on possibility of verbs to form reflexive and/or reciprocal constructions is necessary for rule-based generation of all well formed verb constructions and thus should be integrated in a lexicon.

For example, in *VerbNet*[2] (Kipper et al., 2006), reflexivity is captured with a limited number of English verb classes (5 classes in total) as both syntactic and selectional restrictions on thematic roles. Further, despite being based on Levin's classification of verbs – within which reciprocity is included as one type of the alternations (Levin, 1993) – *VerbNet* does not explicitly distinguish between reciprocal structures and non-reciprocal ones.

In *PropBank* [3] (Palmer et al., 2005), another important lexical resource for English, neither reflexivity nor reciprocity are explicitly described.

As an exception for English, *FrameNet*[4] (Ruppenhofer et al., 2006) introduces the information on reciprocity in the form of the non-lexical semantic frame Reciprocality. However, a systematic way for deriving reciprocal structures is not provided. *FrameNet* does not contain any information on reflexivity.

Concerning non-English lexical resources, let us mention *LexIt* [5] (Lenci et al., 2012), a large-scale lexical resource providing the automatically derived information on distributional properties of Italian verbs, nouns and adjectives. *LexIt* captures reflexives as part of subcategorization frames, however, their reflexive and reciprocal functions are not explicitly distinguished.

For Spanish, the different functions of reflexives are reflected in *DAELE*, a Spanish learner's dictionary (Rennau and Battaner, 2012). Further, similarly as English FrameNet, Spanish FrameNet[6]http://spanishfn.org/ includes the information on reciprocity as well, linking relevant semantic frames by the relation Inheritace to the Reciprocality frame.

For Polish, Patejuk and Przepiórkowski (2015) propose a unified representation of the reflexive *się* within the LFG framework. The analysis has been supplemented with templates for lexical entries as well as with (simplified) f-structures, illustrating the lexicon-grammar interplay. The analysis has been implemented as a part of a large XLE grammar of Polish.

Further, the information on reflexivity is available in treebanks annotated within the Universal Dependencies scheme; however, as (Marković and Zeman, 2018) show the annotation is not consistent. Reciprocity is not annotated here at all.

Finally, reflexivity and reciprocity are distinguished in the Czech and English data in the treebanks of the Prague Dependency Treebank family (Hajič et al., 2012),[7] (Hajič et al., 2018).[8] This information is provided on the tectogrammatical layer, i.e. the layer of deep syntactic annotation. Reflexivity is encoded by the t-lemma (the tectogrammatical lemma, roughly speaking, the deep syntactic lemma) and by the type of coreference: in reflexive constructions, a node with the t-lemma #PersPron, identifying all personal pronouns, is linked by an arrow representing grammatical coreference with its antecedent. Further, reciprocal constructions contain the t-lemma #Rcp, representing the reflexive personal pronoun or surface syntactically unexpressed valency complementation, from which a link identifying grammatical coreference points to its antecedent (Nedoluzhko et al., 2016).

3. Identification of Reflexivity and Reciprocity in Language Data

The identification of verbs allowing for reflexivity and reciprocity is a necessary step in a thorough description of these language phenomena. However, it is obvious at first sight that verbs occurring in reflexive and reciprocal constructions are semantically and syntactically heterogeneous. As a result, it is impossible to formulate straightforward syntactic and/or semantic criteria for identifying these verbs. To solve this tricky task, we have proposed a semi-automatic method which is applied in two steps. In the first step, using grammatical and lexical constraints on reflexivity and reciprocity, we search reflexive and reciprocal constructions in the language data of several corpora. From these corpus searches, candidate verbs allowing for reflexivity and reciprocity were extracted (Section 3.1. and 3.2.). Their set is then extended with semantically similar verbs, identified by means of word embeddings, i.e., representations of verbs in a continuous vector space depending on the contexts in which they appear (Section 3.3.). In

[2] http://verbs.colorado.edu/verb-index/vn/reference.php

[3] http://propbank.github.io/

[4] http://framenet2.icsi.berkeley.edu

[5] http://lexit.fileli.unipi.it/

[6] \unskip\penalty\@M\vrulewidth\z@height\z@depth\dp,

[7] http://ufal.mff.cuni.cz/pcedt2.0-coref, see also http://hdl.handle.net/11234/1-1664

[8] http://ufal.mff.cuni.cz/pdt3.5, see also http://hdl.handle.net/11234/1-2621

both steps, only verbs described in the *VALLEX* lexicon are taken into account.

3.1. Reflexivity

Reflexivity can be applied to verbs that express those events that can be reflected on their initiators. In reflexive constructions, two valency complementations of such verbs have the same referent: one is occupied by a *RM* while the second one, expressed in the subject position, is filled with the antecedent of the *RM*.

Due to the high ambiguity of *RM*s in Czech, see e.g., (Medová, 2009; Svoboda, 2014), reflexive constructions cannot be searched solely on the basis of *RM*s themselves. To identify the reflexive function of *RM*s, we thus took two features into account in our search:

1. *RM*s in reflexive constructions represent the personal pronoun, filling one valency position of a verb. As a result, the form of a *RM* depends on the given valency position, see Table 1.

2. Reflexivity in Czech, as in other Slavic languages, is often marked by the intensifier *sám* 'oneself' that makes it possible to automatically disambiguate it.

The experiments with reflexivity were carried out on a collection of three corpora, namely the SYN4 corpus[9] provided by the Czech National Corpus (Křen et al., 2016; Hnátková et al., 2014), the CzEng 1.0 corpus[10] (Bojar et al., 2011; Bojar et al., 2012), and a large in-house web-corpus. The total size of these data is roughly 435 million sentences (6.8 billion tokens).

Within these corpora, simultaneous presence of features (1) and (2) has been chosen as signaling reflexivity in our corpus search. The corpus data have been searched for occurrences of the *RM*s on either side of the lemma *sám* 'oneself'; *RM*s could have the morphemic form of any case (including prepositional ones), the intensifier *sám* 'oneself' could be either in the nominative or in the same case as the *RM*.

In the context of a pair of a *RM* and the intensifier placed next to each other, a window of at most seven words to either side, but not crossing any conjunctions or punctuation, was extracted. The rightmost verb in this window was determined as the verb forming the reflexive construction.

For example, sentences (7)-(9) were correctly identified as reflexive constructions. In examples (7) and (8), the verb *poznat* 'to get to know', being the rightmost verb, was identified as the one forming the reflexive constructions (in (8), neither the verb *volit* 'to choose' nor the verb *předpokládat* 'assume'). Similarly, in (9), the verb *pomoct* 'to help' (not *dokázat* 'to manage') was detected as the one forming the reflexive construction.

Further, as the SYN4 corpus contains an automatic annotation of phrasemes, we could refine the query by excluding instances of the phrasemes containing the lemma *sám* that do not indicate reflexivity in that portion of the data.

(7) *Poznej sám sebe!*
know oneself$_{acc}$ RM$_{acc}$
'Know yourself!'

(8) *Volit dobro předpokládá nejprve poznat sebe sama.*
to choose goodness assumes first to know RM$_{acc}$ oneself$_{acc}$
'Choosing the good assumes knowing oneself first.'

(9) *Mám strach, že si sama nedokážeš pomoct.*
I have fear that RM$_{dat}$ oneself$_{nom}$ manage help
'I am afraid that you are not able to help yourself.'

A sample of 1,000 candidate reflexive constructions from the SYN4 corpus were subject to a manual analysis. This analysis revealed that for the long forms of *RM*s, *sebe*, *sobě*, *sebou* (be it preceded by a preposition, or not), the selected combination of a *RM* and the intensifier *sám* 'oneself' is a sufficiently reliable clue for the identification of reflexivity, see Table 1.

On the other hand, the search based on the combination of the clitic *RM*s *se* and *si* and the intensifier *sám* 'oneself' did not produce satisfactory results. In order to reduce the negative impact of the ambiguity of the clitic *RM*s, we limited the analysed sample to those verbs that appeared in the SYN4 corpus at least 5 times with the corresponding long forms of the *RM*s. In this case, the combination of the clitic *RM si* with the intensifier *sám* 'oneself' indicated reflexivity in less than 70% of cases, and in the case of the clitic *RM se*, the success rate did not even reach 30% of the sample, see Table 1.

The manual analysis further showed that only 88 of these 1,000 analyzed sentences contained more than one verb. In 85 cases out of them (96%), the heuristic of assigning reflexivity to the rightmost verb was correct.

Type of *RM*	form of *RM*	number	%
$RM_{acc:long}$	*sebe*	199	99
$RM_{dat:long}$	*sobě*	194	97
RM_{long}	$sebe_{gen}$, $sebou_{instr}$	167	84
	prep+$sebe_{gen/acc}$/$sobě_{dat}$/$sebou_{instr}$		
$RM_{acc:clitic}$	*se*	57	29
$RM_{dat:clitic}$	*si*	132	66

Table 1: The manual evaluation of 1,000 candidate reflexive constructions (200 instances for each form of *RM*s) identified by the corpus query in the SYN4 corpus. The numbers represent the share of correctly identified reflexive structures.

Due to the unsatisfactory results for the clitic *RM*s, only instances containing the long forms of the *RM*s were further processed in the follow-up second step. In these instances, 2,699 verb lemmas in total were detected (counting only lemmas that are contained in the *VALLEX* lexicon). Further, the aspectual counterparts of the detected verb lemmas were clustered to 1,792 lexemes as they are implemented in this lexicon, see Section 4.. For the overall statistics, see Table 2.

[9] https://kontext.korpus.cz/first_form?corpname=syn_v4, see also http://hdl.handle.net/11234/1-1846

[10] http://ufal.mff.cuni.cz/czeng/czeng10, see also http://hdl.handle.net/11234/1-1458

Type of *RM*	form of *RM*	lemmas	lexemes
$RM_{\text{acc:long}}$	*sebe*	1,756	1,263
$RM_{\text{dat:long}}$	*sobě*	1,122	861
RM_{long}	$sebe_{\text{gen}}$, $sebou_{\text{instr}}$	2,428	1,680
	prep+$sebe_{\text{gen/acc}}$/$sobě_{\text{dat}}$/$sebou_{\text{instr}}$		
total		2,699	1,792

Table 2: Verbs matching the lexical and grammatical constraints for reflexivity in at least one instance in the investigated corpus data (step 1).

Type of *RM*	form of *RM*	lexemes
$RM_{\text{acc:long}}$	*sebe*	438
$RM_{\text{dat:long}}$	*sobě*	183
RM_{long}	$sebe_{\text{gen}}$, $sebou_{\text{instr}}$	635
	prep+$sebe_{\text{gen/acc}}$/$sobě_{\text{dat}}$/$sebou_{\text{instr}}$	
total		973

Table 3: Candidate lexemes, identified on the basis of lexical and grammatical constraints; these are the lexemes that match the conditions of the corpus query in at least 20 instances (step 1).

Reflexivity – results of step 1: In order to select a reliable portion of the data for further manual processing and adding to the *VALLEX* lexicon, see Section 4., we set a limit of 20 occurrences of the combination of a verb, a *RM* and the intensifier *sám* 'oneself'. Less than a half of the lexemes matching the corpus query satisfy this stricter condition, see Table 3. Further, these lexemes were used in the second step searching for semantically similar verbs as well, see Section 3.3..

3.2. Reciprocity

Reciprocity is applicable to verbs denoting events that can be conceived as mutual. However, reciprocal constructions – similarly to reflexive ones (Section 3.1.) – cannot be identified solely on the basis of *RM*s. The identification of reciprocal constructions had to be grounded in the following features:

1. Verbs appear in reciprocal constructions in plural forms.
2. Reciprocity is signalled by one of the following means:
   (a) *RM*s serve as the primary marker of reciprocity in Czech. In this case, *RM*s represent – as in the case of reflexivity – the personal pronoun filling one valency position of a verb. The form of an individual *RM* thus changes depending on the given valency position.
   (b) The adverbials *navzájem*, *vzájemně*, and *mezi sebou* 'mutually'.[11]
   (c) The bipartite expression *jeden – druhý* 'each other'.

   The adverbials listed in (2b) can be combined with a *RM* or with the expression *jeden – druhý* 'each other'.

For technical reasons, the experiments with reciprocity were carried out only with the data of the SYN4 corpus (275 million sentences, 4.3 billion tokens).
In the corpus search, we took into account only combinations (1)+(2b) and (1)+(2c). Feature (2a) was not specifically postulated in the corpus queries due to the high ambiguity of *RM*s (although *RM*s are in most cases present in the extracted reciprocal constructions).
Two queries for each of the language means expressing reciprocity, *navzájem*, *vzájemně*, *mezi sebou*, and *jeden – druhý*, were formulated (8 in total): one for finite verbs in the plural and one for verbs in the infinitive preceded by a finite verb in the plural. The latter query aimed at reciprocal constructions with modal and auxiliary verbs in which the reciprocal construction should be annotated on the full verb, i.e. the verb in the infinitive.
For example, the reciprocal construction in (10) has been identified on the basis of the pair of features (1)+(2b) (with the adverbial *mezi sebou* 'mutually'). Examples (11) and (12) match the combination of features (1)+(2b) (with the adverbial *vzájemně* and *navzájem* 'mutually', respectively). The reciprocal construction (13) with the expression *jeden – druhý* exemplifies a reciprocal construction matching the combination of features (1)+(2c).

(10) *Gangy se mezi sebou informují.*
gangs RM_{acc} among RM_{instr} inform
'The gangs inform each other.'

(11) … *úkazy elektrické a magnetické vzájemně*
… phenomena electric and magnetic mutually
nesouvisejí …
not relate …
'… electric and magnetic phenomena are not related to each other …'

(12) … *bratři se ale také mohli zabít navzájem.*
… brothers RM_{acc} but also could kill mutually
'… however, brothers also could have killed each other.'

(13) *Podezírali jeden druhého z předpojatosti* …
suspected each other of prejudice …
'They suspected each other of prejudice …'

A sample of 800 found instances (200 for each language means encoding reciprocity) were manually analyzed. This analysis showed that the adverbials *vzájemně*, *navzájem*,

[11] Although the expression *mezi sebou* consists of the preposition *mezi* 'among' and the long reflexive *sebou*, it is not licensed by valency of verbs forming reciprocal constructions. The reciprocal construction typically contains yet another *RM* that fills a valency position of the verb and that thus serves as the primary marker of reciprocity while the prepositional group *mezi sebou* has the same function as the adverbs *navzájem* or *vzájemně* 'mutually'. For example in (10), it is the clitic *RM se* (standing for the reflexive pronoun in the accusative case) that occupies one valency position of the verb *informovat* 'to inform', serving thus as the primary marker of reciprocity.

and *mezi sebou* 'mutually' are reliable indicators of reciprocity. In contrast, the bipartite expression *jeden – druhý* 'each other' – in addition to encoding reciprocity – is part of various adverbials expressing the manner of the actions rather than reciprocity, see example (14). As a result, the search based on this feature gave much less satisfactory results. For the manual evaluation of the samples, see Table 4.

(14) *Dny plynuly jeden za druhým.*
days passed one behind second
'Days passed one by one.'

Reciprocal marker	number	%
vzájemně 'mutually'	179	90
navzájem 'mutually'	179	90
mezi sebou 'mutually'	162	81
jeden – druhý 'each other'	95	48

Table 4: The manual analysis of 800 candidate reciprocal constructions (200 instances for each reciprocal marker) identified by the corpus query. The numbers represent the share of correctly identified reciprocal structures.

Due to the unsatisfactory results for the expression *jeden – druhý* 'each other', see Table 4, only the results of the corpus queries with the adverbials *navzájem*, *vzájemně*, and *mezi sebou* 'mutually' were further processed in the follow-up second step. On their basis, 2,245 verb lemmas were detected that form 1,535 lexemes (counting only those contained in the *VALLEX* lexicon). For the overall statistics, see Table 5.

Reciprocal marker	lemmas	lexemes
navzájem 'mutually'	1,884	1,345
vzájemně 'mutually'	1,674	1,228
mezi sebou 'mutually'	1,248	951
total	2,245	1,535

Table 5: Total number of *VALLEX* verbs matching the lexical and grammatical constraints in at least one occurrence in the SYN4 corpus.

Reciprocal marker	lexemes
navzájem 'mutually'	363
vzájemně 'mutually'	294
mezi sebou 'mutually'	222
total	605

Table 6: Candidate verbs for reciprocal constructions identified on the basis of lexical and grammatical constraints; these are the *VALLEX* lexemes with at least 20 instances matching the conditions of the corpus query (step 1).

Reciprocity – results of step 1: Similarly as for reflexivity, we set the limit of 20 occurrences of the combination of a verb and one of the adverbials *navzájem*, *vzájemně*, and *mezi sebou* 'mutually' to select clear candidates for further manual processing and adding to the *VALLEX* lexicon, see Section 4.. Two fifths of the found lexemes satisfied this criterion, see Table 6. These lexemes were also used in the word2vec model, see Section 3.3..

3.3. Searching for Semantically Similar Verbs with word2vec

To expand the list of verbs allowing for reflexive and/or reciprocal constructions obtained by the manually tuned corpus search (step 1, Sections 3.1. and 3.2.), we used word2vec (Mikolov et al., 2013). It was trained on the same lemmatized corpora on which we carried out step 1: the Czech National Corpus (the SYN4 corpus), the CzEng 1.0 corpus, and an in-house web-corpus in the case of reflexivity, and just the SYN4 corpus in the case of reciprocity. In these data, verb lemmas were substituted by the respective verb lexemes.

Reflexivity. Three models were trained for reflexivity using the gensim library[12] (Řehůřek and Sojka, 2010) with the following parameters: vector size 512, context window 5, skip-gram training algorithm and negative sampling. The minimum word count was set to 20 as the algorithm can hardly learn properties of scarcely seen words.

M_{RM} In the training data for this model, candidate verb lexemes were marked by the suffix *_R* in all constructions identified in the first steps. (As a result, each lexeme that occurred in reflexive constructions has two representations in this model – one with and one without the suffix *_R*.)

M_{R-M} In this model, the intensifier *sám* 'oneself' serving as a reflexivity marker was erased from the training data, while maintaining the marking of the identified lexemes by the suffix *_R*. This model thus still provides two representations for each lexeme.[13]

M_{B} The input data for this baseline model were not adjusted at all. The purpose of this model was to verify whether adding the information on reflexivity of verbs (the suffix *_R*) helps the algorithm to better identify new reflexive constructions or not.

In all three models, we investigated an immediate neighborhood of those lexemes that had at least 20 instances of potentially reflexive constructions found in step 1, see Table 3 providing the statistics on these lexemes.[14] For each of these lexemes, we listed those lexemes from the *VALLEX* lexicon that were among their 20 closest neighbors with cosine similarity at least 0.5.

[12] `http://radimrehurek.com/gensim`

[13] It appeared that in model M_{RM} the presence of the intensifier *sám* 'oneself', which occurred in the vicinity of all lexemes marked by the *_R* suffix, is a very strong indicator, leading to an undesirable split between the representations of lexemes with this suffix and those without it, see Figure 1. Model M_{R-M} thus excluded the intensifier *sám* 'oneself', preserving only the reflexive marking of the respective lexemes.

[14] In addition to the 973 *VALLEX* lexemes, 175 verb lexemes not contained in *VALLEX* but found in step 1 in at least 20 potentially reflexive constructions were also used in this step. We expected that in the neighborhood of these 175 lexemes other candidates for reflexivity contained in *VALLEX* could occur as well.

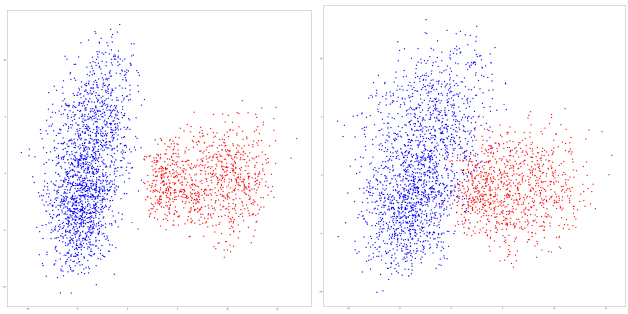

Figure 1: Projection of the representation of *VALLEX* lexemes in models M_{RM} (on the left) and M_{R-M} (on the right) onto two principal components (as determined by the PCA algorithm, using the `scikit-learn` library (Pedregosa et al., 2011)). Points corresponding to lexemes with the *_R* suffix are in red, those without the suffix are in blue.

From this output list of lexemes we were interested in only those that have not been identified as candidates for manual annotation of reflexivity yet. (i.e., in those that were not identified in step 1 at all or in those that were found in less than 20 potentially reflexive constructions). See Table 7 for the numbers of newly identified candidate lexemes.

lexemes in step 1	M_{R-M}	M_{RM}	M_B	$M_{RM} \cup M_{R-M}$
not identified	758	904	79	908
frequency < 20	1	0	746	1
total	759	904	825	909

Table 7: New candidate lexemes for reflexivity identified using word2vec.

A random sample of 100 verb lexemes for each of the above described three models were subject to a manual analysis. The analysis revealed that the information on reflexivity provided in models M_{RM} and M_{R-M} produced satisfactory results, in contrast to model M_B, see Table 8. These results corroborate the hypothesis that the reflexivity marking present in M_{RM} and M_{R-M} greatly assists in detecting verb candidates for reflexivity.

	number (=%)
M_{RM}	83
M_{R-M}	88
M_B	34

Table 8: Results of manual evaluation of a random sample of 100 candidate lexemes for reflexivity for each of the models.

Reflexivity – results of step 2: As a result, 909 verb lexemes were identified by the M_{RM} and M_{R-M} models as candidates for further manual processing and adding to the VALLEX lexicon, see Section 4.. The fact that virtually no verb lexeme with low number of candidate reflexive constructions identified in step 1 appears among the new candidates supports the adequacy of the frequency limit (at least 20 instances matching the query) set in Section 3.1..

Reciprocity. In case of reciprocity, the procedure in step 2 was much the same as in case of reflexivity. We trained three models, two specific for reciprocity and one identical for reciprocity and reflexivity: model M_{RM} on the data with the lexemes in candidate reciprocal constructions identified in step 1 marked by the suffix *_R*, M_{R-M} on the same data but with the adverbials *vzájemně*, *navzájem* and *mezi sebou* (reciprocity markers) erased, and the baseline model M_B on data with no information on reciprocity.
Surprisingly, model M_{RM} did not produce any new candidates: in the investigated neighborhoods, it found only lexemes already identified in step 1 as reliable candidates for reciprocity. As to model M_{R-M}, just two new candidate lexemes were found, namely *tázat se* and *otázat se* 'to ask'.[15] Finally, model M_B – similarly to the experiment with reflexivity – produced new candidates (778 of them), but with poor reliability as revealed by a manual evaluation (only 26 out of 100 manually investigated candidates actually form reciprocal constructions).

4. Reflexivity and Reciprocity in *VALLEX*

For the representation of language phenomena at the interface between semantics and syntax, the valency lexicon *VALLEX*[16] (Kettnerová et al., 2012) takes advantage of a division into the data component (Section 4.1.) and the grammar component (Section 4.2.). In case of reflexivity and reciprocity, the annotation in the data component was carried out primarily for the candidates obtained in steps 1 and 2 of the experiment (Section 3.); the rules in the grammar component were then formulated on the basis of the corpus samples gathered in step 1 and exploited as examples in the data component.

4.1. Data Component

The data component represents an inventory of Czech verb lexemes, where a lexeme associates a set of verb lemmas with a set of lexical units (corresponding to the individual meanings of verbs). Lexical units are characterized by valency frames underlying their deep syntactic structures. These frames are enriched with possible morphemic forms of valency complementations, indicating their surface syntactic expression. Valency frames describe unmarked constructions of lexical units of verbs, i.e., active, non-reflexive, and non-reciprocal constructions. If it is relevant, each lexical unit of a verb carries information on the possibility to create marked syntactic constructions, i.e., reflexive and reciprocal constructions.[17] Pair(s) of those valency complementations that can be involved in reflexivity and/or reciprocity are identified in the relevant attributes, see Figure 2.[18]

[15] Both these verbs are inherently reflexive verbs, i.e. the reflexive *se* represents a part of verb lemma (not the reflexive pronoun).

[16] `http://ufal.mff.cuni.cz/vallex`, see also `http://hdl.handle.net/11234/1-2307`

[17] Other listed constructions, such as the passive, are left aside here.

[18] The functions of various types of *RM*s and their representation in the valency lexicon VALLEX is thoroughly described in (Kettnerová and Lopatková, 2019).

myslet / myslit	
	≈ *brát zřetel* 'to take into account'
frame	$\mathsf{ACT}^{\mathsf{obl}}{}_{1}$ $\mathsf{PAT}^{\mathsf{obl}}{}_{na+\mathsf{acc,dcc}}$
example	mysleli při stavbě na hendikepované 'during the construction, they took into account people with disabilities'
class	mental action
reflex	ACT-PAT
recipr	ACT-PAT

Figure 2: Sample lexical unit for the verb lexeme represented by the lemma variants *myslet* and *myslit*.

Information on reflexivity and reciprocity will be manually added to relevant lexical units of reliable candidate lexemes resulting from the experiment described in Section 3.. As a result, information on reflexivity will be recorded with lexical units of 1.882 lexemes (973 lexemes suggested in step 1 and further 909 coming from step 2). Information on reciprocity will be supplied with lexical units of 607 lexemes (605 lexemes identified in step 1 and additional two lexemes in step 2).

4.2. Grammar Component

Reciprocity and reflexivity bring about systemic changes in surface syntactic structure of verbs that can be described by grammatical rules. In *VALLEX*, these rules are stored in the grammar component, see (Lopatková et al., 2016; Lopatková and Kettnerová, 2016; Kettnerová and Lopatková, 2018) for more detailed description. A simplified rule describing reflexivity in Czech is given in Table 3.

Reflexivity affecting X and Y valency complementations		
reflex	Xnom & Y	form of *RM*
forms of Y	gen → $RM_{\mathsf{gen:long}}$	*sebe*
	dat → $RM_{\mathsf{dat:clitic/long}}$	*si*, *sobě*
	acc → $RM_{\mathsf{acc:clitic/long}}$	*se*, *sebe*
	instr → $RM_{\mathsf{instr:long}}$	*sebou*
	prep+case → prep+RM_{long}	$sebe_{\mathsf{gen/acc}}$/ /$sobě_{\mathsf{dat}}$/$sebou_{\mathsf{instr}}$
	other forms → Ø	Ø
oblig.	Y	

Figure 3: Simplified rule for deriving reflexive constructions. As *RM*s in reflexive constructions stand for the reflexive personal pronoun, they can take long as well as (in dative and accusative) clitic forms, depending on their position in a sentence.

For example, the lexical unit of the verb *myslet* 'to think' with the meaning 'to think of, take somebody/something into account, consider' has the valency frame provided in (15a), see example (15b). When the rule in Table 3 is applied to the valency frame of this verb, the valency frame underlying the usage of this lexical unit in reflexive constructions is derived, see the frame in (16a).

(15) a. $\mathsf{ACT}_{\mathsf{nom}}$ $\mathsf{PAT}_{\mathsf{na+acc,dcc}}$

b. *Architekti mysleli při stavbě na vozíčkáře.*
'When constructing achitects took wheelchair users into account.'

(16) a. $\mathsf{ACT}_{\mathsf{nom}}$ $\mathsf{PAT}_{\mathsf{na+RMacc:long}}$

b. *Jan myslel jen na sebe.*
'John took only himself into account.'

5. Conclusion

We have proposed a method for the identification of verbs allowing for reflexivity and reciprocity, consisting of two steps. The first step is relying on grammatical and lexical constraints and their application in corpus queries. The follow-up step makes use of word embeddings for detecting verbs semantically similar to those that were detected in the first step. Manual annotation of samples of the candidates from steps 1 and 2 suggests that in case of reflexivity, both steps provide sufficiently reliable candidates. In case of reciprocity, satisfactory results were obtained only in the first step; the second step merely confirms that lexemes that regularly form reciprocal constructions are typically semantically related to each other.
The resulting candidate verbs will be further used in the manual annotation of reflexivity and reciprocity in the *VALLEX* lexicon.

6. Acknowledgements

The research reported in this paper has been supported by the project No. 18-03984S, *Between Reciprocity and Reflexivity: The Case of Czech Reciprocal Constructions*, of the Czech Science Foundation (GAČR) and partially also by the project No. LM2015071, *LINDAT/CLARIN*, of the Ministry of Education, Youth and Sports of the Czech Republic.
This work has been using language resources stored and distributed by the project No. LM2015071, *LINDAT-CLARIN*, of the Ministry of Education, Youth and Sports of the Czech Republic.

7. Bibliographical References

Bojar, O., Žabokrtský, Z., Dušek, O., Galuščáková, P., Majliš, M., Mareček, D., Maršík, J., Novák, M., Popel, M., and Tamchyna, A. (2011). Czech-English Parallel Corpus 1.0 (CzEng 1.0). LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Bojar, O., Žabokrtský, Z., Dušek, O., Galuščáková, P., Majliš, M., Mareček, D., Maršík, J., Novák, M., Popel, M., and Tamchyna, A. (2012). The Joy of Parallelism with CzEng 1.0. In *Proceedings of LREC2012*, Istanbul, Turkey, May. ELRA, European Language Resources Association.

Evans, N., Gaby, A., and Nordlinger, R. (2007). Valency mismatches and the coding of reciprocity in Australian languages. *Linguistic Typology*, 11:541–597.

Frajzyngier, Z. and Walker, T., editors. (2000a). *Reciprocals. Forms and Functions*, volume 41 of *Typological Studies in Language*. John Benjamins, AmsterdamPhiladelphia.

Frajzyngier, Z. and Walker, T., editors. (2000b). *Reflexives. Forms and Functions*, volume 40 of *Typological Studies in Language*. John Benjamins, AmsterdamPhiladelphia.

Geniušienė, E. (1987). *The Typology of Reflexives*. Mouton de Gruyter, Berlin–New York–Amsterdam.

Hajič, J., Hajičová, E., Panevová, J., Sgall, P., Bojar, O., Cinková, S., Fučíková, E., Mikulová, M., Pajas, P., Popelka, J., Semecký, J., Šindlerová, J., Štěpánek, J., Toman, J., Urešová, Z., and Žabokrtský, Z. (2012). Announcing Prague Czech-English Dependency Treebank 2.0. In Calzolari, N. et al., editors, *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 3153–3160, Paris, France. ELRA.

Hajič, J., Bejček, E., Bémová, A., Buráňová, E., Hajičová, E., Havelka, J., Homola, P., Kárník, J., Kettnerová, V., Klyueva, N., Kolářová, V., Kučová, L., Lopatková, M., Mikulová, M., Mírovský, J., Nedoluzhko, A., Pajas, P., Panevová, J., Poláková, L., Rysová, M., Sgall, P., Spoustová, J., Straňák, P., Synková, P., Ševčíková, M., Štěpánek, J., Urešová, Z., Vidová Hladká, B., Zeman, D., Zikánová, Š., and Žabokrtský, Z. (2018). Prague Dependency Treebank 3.5. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Hnátková, M., Ken, M., Procházka, P., and Skoumalová, H. (2014). The SYN-series corpora of written Czech. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 160–164, Reykjavik. ELRA.

Kemmer, S. (1993). *The Middle Voice*. John Benjamins, Amsterdam–Philadelphia.

Kettnerová, V. and Lopatková, M. (2018). Lexicographic Potential of the Syntactic Properties of Verbs: The Case of Reciprocity in Czech. In *XVIII EURALEX International Congress, Lexicography in Global Contexts*, pages 685–698, Ljubljana. Ljubljana University Press, Faculty of Arts.

Kettnerová, V. and Lopatková, M. (2019). Reflexives in Czech from a Dependency Perspective. In Gerdes, K. and Kahane, S., editors, *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, Syntaxfest 2019)*, pages 14–25, Paris, France. Association for Computational Linguistics.

Kettnerová, V., Lopatková, M., and Bejček, E. (2012). The Syntax-Semantics Interface of Czech Verbs in the Valency Lexicon. In *Proceedings of the XV EURALEX International Congress*, pages 434–443, Oslo. University of Oslo.

Kipper, K., Korhonen, A., Ryant, N., and Palmer, M. (2006). A Large-Scale Extension of VerbNet with Novel Verb Classes. In Corino, E., Marello, C., and Onesti, C., editors, *Proceedings of the 12th EURALEX International Congress*, pages 173–184, Torino, Italy. Edizioni dell'Orso.

König, E. and Gast, V., editors. (2008). *Reciprocals and Reflexives: Theoretical and Typological Explorations*. Mouton de Gruyter, Berlin.

Křen, M., Cvrček, V., Čapka, T., Čermáková, A., Hnátková, M., Chlumská, L., Jelínek, T., Kovářík ová, D., Petkevič, V., Procházka, P., Skoumalová, H., Škrabal, M., Truneček, P., Vondřička, P., and Zasina, A. (2016). SYN v4: large corpus of written Czech. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Lenci, A., Lapesa, G., and Bonansinga, G. (2012). LexIt: A Computational Resource on Italian Argument Structure. In Calzolari, N. et al., editors, *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'12)*, Paris, France, May. ELRA.

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

Lopatková, M. and Kettnerová, V. (2016). Alternations: From Lexicon to Grammar And Back Again. In Hajičová, E. and Boguslavsky, I., editors, *Proceedings of the Workshop on Grammar and Lexicon: Interactions and Interfaces (GramLex)*, pages 18–27, Ōsaka, Japan. ICCL, The COLING 2016 Organizing Committee.

Lopatková, M., Kettnerová, V., Bejček, E., Vernerová, A., and Žabokrtský, Z. (2016). *Valenční slovník českých sloves VALLEX*. Karolinum, Praha.

Marković, S. and Zeman, D. (2018). Reflexives in Universal Dependencies. In *Proceedings of the 17th International Workshop on Treebanks and Linguistic Theories (TLT 2018)*, pages 131–146, Linköping, Sweden. Linköping University Electronic Press.

Medová, L. (2009). *Reflexive Clitics in the Slavic and Romance Languages. A Comparative View from an Antipassive Perspective*. Ph.D. thesis, Princeton University, Princeton, NJ, USA.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., and Dean, J. (2013). Distributed Representations of Words and Phrases and Their Compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'13, pages 3111–3119, USA.

Nedjalkov, V. P., editor. (2007). *Reciprocal Constructions*. John Benjamins, Amsterdam–Philadelphia.

Nedoluzhko, A., Novák, M., Cinková, S., Mikulová, M., and Mírovský, J. (2016). Coreference in Prague Czech-English Dependency Treebank. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 23–28, Paris, France. ELRA.

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

Patejuk, A. and Przepiórkowski, A. (2015). An LFG Analysis of the So-called Reflexive Marker in Polish. In Butt, M. and King, T. H., editors, *Proceedings of the LFG15 Conference*, CSLI Publications, pages 270–288.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E.

(2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May. ELRA.

Renau, I. and Battaner, P. (2012). Using CPA to represent Spanish pronominal verbs in a learners dictionary. In Fjeld, R. V. and Torjusen, J. M., editors, *Proceedings of the 15th EURALEX International Congress*, pages 350–361, Oslo, Norway, aug. Department of Linguistics and Scandinavian Studies, University of Oslo.

Ruppenhofer, J., Ellsworth, M., Petruck, M. R. L., Johnson, C. R., and Scheffczyk, J. (2006). *FrameNet II: Extended Theory and Practice*. International Computer Science Institute, California, Berkeley.

Svoboda, O. (2014). Functions of the Czech reflexive marker *se/si*. Master's thesis, Universitet Leiden.