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; Frayzynger and Walker, 2000b; Frayzynger and Walker, 2000a). These language phenomena are typically encoded by reflexive markers (henceforth RM). However, cross-linguistic studies evidence that RMs typically serve various functions in a language, crossing boundaries of reflexivity and reciprocity, see esp. (Genušiene, 1987) and (Kemmer, 1993). In Czech, which is the focus of this work, RMs are involved in various constructions. In addition to reflexive (1) and reciprocal constructions (2), RMs encode impersonal passive constructions (3), middle constructions (4), anticausative constructions (5) or constructions with inherently reflexive verbs (6).1

(1) Petr myslej na sebe.
Peter thought only of RM
‘Peter thought only of himself.’

(2) Rodíte si (vzájemně) pomáhali.
parents RM (mutually) helped
‘Parents helped each other.’

(3) Z České republiky se elektřina vyváží.
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 rozbito.
window RM broke
‘The window broke.’

(6) Petr se smál od srdce.
Peter RM laughed heartily
‘Peter laughed heartily.’

The high ambiguity of RMs complicates their description, often leading different theories to contradictory conclusions. Moreover, it produces many inconsistencies in their annotation, see (Markovic 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 RMs are thus highly beneficial for both theoretical linguistic research and NLP applications. Moreover, high frequency of RMs 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 RMs 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 constructions in Czech, it is evident that there is a need for semi-automatic methods for the detection of such constructions.

1Similar functions are exhibited by RMs in other Slavic and Romance languages as well (Medova, 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 RMs 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\(^4\) (Kipper et al., 2006), reflexivity is captured with a limited number of English verb classes (5 classes in total) as both syntactic and selective 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’s classification of verbs – within which reciprocity is captured with a limited number of English verb classes (5 classes in total) as both syntactic and selective 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\(^5\) (Palmer et al., 2005), another important lexical resource for English, neither reflexivity nor reciprocity are explicitly described.

As an exception for English, FrameNet\(^6\) (Ruppenhofer et al., 2006) introduces the information on reciprocity in the form of the non-lexical semantic frame Reciprocity. 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\(^7\) (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 (\(\text{\textmd{RC3}}\), see also http://verbs.colorado.edu/verb-index/vn/reference.php http://propbank.github.io/ http://framenet2.icsi.berkeley.edu http://litect.tildei.unipi.it/). Reflexivity and reciprocity are distinguished in the Czech and English data in the treebanks of the Prague Dependency Treebank family (Hajiˇc et al., 2012). 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., 2010).

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

1. RMs 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 |.

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| provided by the Czech National Corpus (Kren et al., 2016; Hnatková et al., 2014), the CzEng 1.0 corpus| (Bojar et al., 2011; Bojar et al., 2012), and a large in-house webcorpus. The total size of these data is roughly 435 million sentences (6.8 billion tokens).

Within these corpora, simultaneous presence of features | and | has been chosen as signaling reflexivity in our corpus search. The corpus data have been searched for occurrences of the RMs on either side of the lemma sám ‘oneself’; RMs 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 | and | were correctly identified as reflexive constructions. In examples | and |, the verb poznat ‘to get to know’, being the rightmost verb, was identified as the one forming the reflexive constructions (in |, neither the verb volit ‘to choose’ nor the verb předpokládat ‘to assume’). Similarly, in |, the verb pomoci ‘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.

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 RMs, 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 |.

On the other hand, the search based on the combination of the clitic RMs 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 RMs, 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 RMs. 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 |.

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.

<table>
<thead>
<tr>
<th>Type of RM</th>
<th>form of RM</th>
<th>number</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>RM_{acc, long}</td>
<td>sebe</td>
<td>199</td>
<td>99</td>
</tr>
<tr>
<td>RM_{dat, long}</td>
<td>sobě</td>
<td>194</td>
<td>97</td>
</tr>
<tr>
<td>RM_{long}</td>
<td>sebou, sebo, sobu</td>
<td>167</td>
<td>84</td>
</tr>
<tr>
<td>RM_{acc, clitic}</td>
<td>se</td>
<td>57</td>
<td>29</td>
</tr>
<tr>
<td>RM_{dat, clitic}</td>
<td>si</td>
<td>132</td>
<td>66</td>
</tr>
</tbody>
</table>

Table 1: The manual evaluation of 1,000 candidate reflexive constructions (200 instances for each form of RMs) 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 RMs, only instances containing the long forms of the RMs 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 |.

For the overall statistics, see Table |.
Reciprocity – 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 3.1.1 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 3.1.1) – cannot be identified solely on the basis of RMs. 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) RMs serve as the primary marker of reciprocity in Czech. In this case, RMs 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’.

   (c) The bipartite expression *jeden – druhý* ‘each other’.

The adverbials listed in (b) 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 (11)+ (13) and (11)+(12). Feature (12) was not specifically postulated in the corpus queries due to the high ambiguity of RMs (although RMs 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 (11) has been identified on the basis of the pair of features (11)+(12) (with the adverbial *mezi sebou* ‘mutually’). Examples (11) and (13) match the combination of features (11)+(12) (with the adverbial *vzájemně* and *navzájem* ‘mutually’, respectively). The reciprocal construction (11) with the expression *jeden – druhý* exemplifies a reciprocal construction matching the combination of features (11)+(13).

(10) **Gangy se mezi sebou informují.**

gangs RM}_{sec} among RM}_{pred} inform

‘The gangs inform each other.’

(11) **… úkazy elektrické a magnetické vzájemně …** phenomena electric and magnetic mutually

… nesouvisel …

… electric and magnetic phenomena are not related to each other …

(12) **… bratři se ale také mohli zabít navzájem.**

… brothers RM}_{sec} but also could kill mutually

… however, brothers also could have killed each other.

(13) **Podezírali jeden druhého z předpovjatosti …**

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*, 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 (11), 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.

<table>
<thead>
<tr>
<th>Type of RM</th>
<th>form of RM</th>
<th>lemmas</th>
<th>lexemes</th>
</tr>
</thead>
<tbody>
<tr>
<td>RM}_{sec}</td>
<td>sebe</td>
<td>1,756</td>
<td>1,263</td>
</tr>
<tr>
<td>RM}_{pred}</td>
<td>sobě</td>
<td>1,122</td>
<td>861</td>
</tr>
<tr>
<td>RM_{long}</td>
<td>sebe prep, sebou</td>
<td>2,428</td>
<td>1,680</td>
</tr>
<tr>
<td></td>
<td>prep+sebe prep,</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>sobě prep, sobou</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Type of RM</th>
<th>form of RM</th>
<th>lemmas</th>
<th>lexemes</th>
</tr>
</thead>
<tbody>
<tr>
<td>RM}_{sec}</td>
<td>sebe</td>
<td>438</td>
<td></td>
</tr>
<tr>
<td>RM}_{pred}</td>
<td>sobě</td>
<td>183</td>
<td></td>
</tr>
<tr>
<td>RM_{long}</td>
<td>sebe prep, sebou</td>
<td>635</td>
<td></td>
</tr>
<tr>
<td></td>
<td>prep+sebe prep,</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>sobě prep, sobou</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Type of RM</th>
<th>form of RM</th>
<th>lemmas</th>
<th>lexemes</th>
</tr>
</thead>
<tbody>
<tr>
<td>RM}_{sec}</td>
<td>sebe</td>
<td>1,122</td>
<td>861</td>
</tr>
<tr>
<td>RM}_{pred}</td>
<td>sobě</td>
<td>438</td>
<td></td>
</tr>
<tr>
<td>RM_{long}</td>
<td>sebe prep, sebou</td>
<td>635</td>
<td></td>
</tr>
<tr>
<td></td>
<td>prep+sebe prep,</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>sobě prep, sobou</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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

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).
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 6.

(14) Dny plyny jedn za druhým.,
days passed one behind second
‘Days passed one by one.’

### Table 4: 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.

<table>
<thead>
<tr>
<th>Reciprocal marker</th>
<th>number</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>vzájemně ‘mutually’</td>
<td>179</td>
<td>90</td>
</tr>
<tr>
<td>navzájem ‘mutually’</td>
<td>179</td>
<td>90</td>
</tr>
<tr>
<td>mezi sebou ‘mutually’</td>
<td>162</td>
<td>81</td>
</tr>
<tr>
<td>jeden – druhý ‘each other’</td>
<td>95</td>
<td>48</td>
</tr>
</tbody>
</table>

Due to the unsatisfactory results for the expression jeden – druhý ‘each other’, see Table 6, only the results of the corpus queries with the adverbials navzájem, vzájemně, and mezi sebou ‘mutually’ were further processed in the follow-up 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 6.

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

<table>
<thead>
<tr>
<th>Reciprocal marker</th>
<th>lemmas</th>
<th>lexemes</th>
</tr>
</thead>
<tbody>
<tr>
<td>navzájem ‘mutually’</td>
<td>1,884</td>
<td>1,345</td>
</tr>
<tr>
<td>vzájemně ‘mutually’</td>
<td>1,674</td>
<td>1,228</td>
</tr>
<tr>
<td>mezi sebou ‘mutually’</td>
<td>1,248</td>
<td>951</td>
</tr>
<tr>
<td>total</td>
<td>2,245</td>
<td>1,535</td>
</tr>
</tbody>
</table>

### 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).

<table>
<thead>
<tr>
<th>Reciprocal marker</th>
<th>lexemes</th>
</tr>
</thead>
<tbody>
<tr>
<td>navzájem ‘mutually’</td>
<td>363</td>
</tr>
<tr>
<td>vzájemně ‘mutually’</td>
<td>294</td>
</tr>
<tr>
<td>mezi sebou ‘mutually’</td>
<td>222</td>
</tr>
<tr>
<td>total</td>
<td>605</td>
</tr>
</tbody>
</table>

### 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 3.1. 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\(^{12}\) (Rehůrek 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.

- \(M_{\text{ref}}\): 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.)
- \(M_{\text{ref-s}}\): 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.\(^{13}\)
- \(M_{\text{ref-t}}\): 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 6 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.

\[^{12}\text{http://radimrehurek.com/gensim}\]

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

\[^{14}\text{In addition to the 973 VALLEX lexemes, 175 verb lexemes not contained in VALLEX but found in step 1 in at least 20 potentially reflexive constructions were also used in this step. We expected that in the neighborhood of these 175 lexemes other candidates for reflexivity contained in VALLEX could occur as well.}\]
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 \( M_{\text{RM}} \) on the data with the lexemes in candidate reciprocal constructions identified in step 1 marked by the suffix \( _R \), \( M_{\text{RM,S}} \) on the same data but with the adverbials vzájemně, navzájem and mezi sebou (reciprocity markers) erased, and the baseline model \( M_{\theta} \) on data with no information on reciprocity.

Surprisingly, model \( M_{\theta} \) 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 \( M_{\text{RM,S}} \), just two new candidate lexemes were found, namely tázat se and otázat se ‘to ask’. Finally, model \( M_{\text{RM}} \) – 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\(^{13}\) (Kettnerová et al., 2012) takes advantage of a division into the data component (Section 3.1) and the grammar component (Section 3.2). 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. Pair(s) of those valency complementations that can be involved in reflexivity and/or reciprocity are identified in the relevant attributes, see Figure 4.\(^{15}\)

---

\(^{13}\)Both these verbs are inherently reflexive verbs, i.e. the reflexive se represents a part of verb lemma (not the reflexive pronoun).

\(^{14}\)Pair(s) of those valency complementations that can be involved in reflexivity and/or reciprocity are identified in the relevant attributes, see Figure 4.

\(^{15}\)http://ufal.mff.cuni.cz/vallex

\(^{16}\)http://hdl.handle.net/11234/1-2307

\(^{17}\)Other listed constructions, such as the passive, are left aside here.
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. ACT\textsuperscript{rat}, PAT\textsubscript{na-acc-dec}

b. Architekti mysleli při stavbě na vozíčkáře. ‘When constructing architecs took wheelchair users into account.’

(16)  a. ACT\textsuperscript{rat}, PAT\textsubscript{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


