Reflexives in Universal Dependencies

Sonja Marković, Daniel Zeman
Charles University, Faculty of Mathematics and Physics, Prague, Czechia
{markovic,zeman}@ufal.mff.cuni.cz

Abstract
We explore the annotation of reflexives in Universal Dependencies (UD) 2.2 treebanks (Nivre et al., 2018), with a stronger focus on Slavic languages. We have tried to find out if the current guidelines are transparent and clear enough for the annotators to follow them successfully. We point out a number of inconsistencies in the current annotation across languages, and propose improvements—sometimes of the guidelines, but mostly of the annotation. The goal of the paper is to contribute to more consistent annotation of reflexives in future releases of UD, which, in turn, will enable broader cross-linguistic studies of this phenomenon.

Keywords: Universal Dependencies, reflexive pronoun, reflexive construction, annotation consistency.

1 Introduction

Reflexive verbs can be found in a significant number of languages, varying from language to language (Geniušienė, 1987, p. 361). This term can be used for verbs with a reflexive marker **(RM)** as their element (affix, inflection, etc.) or as a part of their environment (particle, pronoun, etc.) (Geniušienė, 1987, p. 237). The term 'reflexive' implies that the main function of the reflexive marker is to mark reflexivity, i.e., that a reflexive marker is coreferential with the subject of the clause in which it appears. Nonetheless, marking reflexivity is not the only function of the reflexive marker; it is just one of many functions (Svoboda, 2014, p. 1). The term 'reflexive' is certainly ambiguous but it is broadly used in literature and we will use it in this paper, too. Reflexive markers can be (Geniušienė, 1987, p. 242 and 303):

- affixal morphemes (e.g. prefixes like *t-* in Amharic; suffixes like *-l-* and *-n-* in Turkic languages or *-sja* in Russian; infixes like *-mi-* in Lingala);
- changes in the verbal paradigm (e.g. a change in the agreement paradigm, or a special reflexive conjugation);
- a word or phrase (refl. pronoun like *zibun* "oneself" in Japanese, or a series of pronouns like the Swedish *mig/dig/oss/er/sig*);
- a more or less desemanticized noun meaning "soul", "head", "body", "self" etc., sometimes with a possessive (Basque *buru* "head", e.g., *nere burua* (lit. *my head*) "myself", *bere burua* (lit. *his head*) "himself").

The feature of reflexivity can also apply to modifiers that are neither reflexive verbs nor their arguments. This is the case with reflexive possessives, which indicate that the modified noun is possessed by, or relates to the subject. The Czech example in Figure 1 shows both a reflexive pronoun *(sebe)* and a reflexive possessive *(svého)*.

Both words are coreferential with the subject of the clause, here *Jana*. Thus the two people registered were Jana and Jana's brother (not someone else's brother). The reflexive pronoun is a noun phrase of its own, while reflexive possessives are typically embedded in larger noun phrases. In this paper we focus on various functions of reflexive markers that replace noun phrases or appear at positions similar to those of noun phrases; possessives are not relevant for us, and we will mostly ignore them from now on.

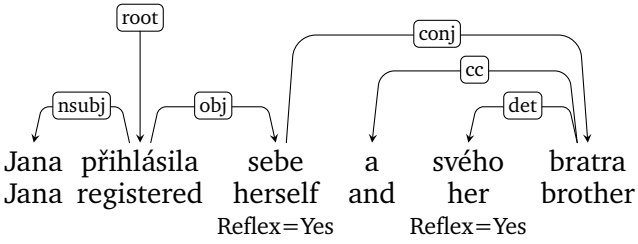

Figure 1: Reflexive object and possessive.

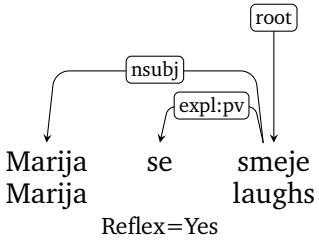

Figure 2: Inherently reflexive.

Furthermore, reflexives in some languages have the same or similar form as intensifiers (emphatic pronouns) such as *himself* in

(1) *John **himself** built this house.*

While in most European languages intensifiers differ from reflexive pronouns (e.g. German *selbst:sich*; Italian *stesso:se*), in many other languages (Turkic, Finno-Ugric, Indic, Persian...), intensifiers are identical with reflexives in form, although not in distribution (König and Siemund, 2000, p. 41). Due to the limited space (and to the language groups focused on in this paper), we do not take intensifiers into account.

Although reflexivity has been intensively studied for the last couple of decades, both in individual languages and from a cross-linguistic perspective, annotating all the functions of all reflexive markers is a complex and sensitive task (Kettnerová and Lopatková, 2014).

In the present paper, we look at reflexives in the context of one particular treebank annotation scheme—Universal Dependencies **(UD)** (Nivre et al., 2016). The nature of the Universal Dependencies project makes annotating reflexive markers particularly difficult, since it aims at developing cross-linguistically consistent treebank annotation for many languages. Discussing the annotation options for reflexives in their various functions is thus very important. This paper is just a starting point for a more comprehensive study of reflexives in UD. At present, such a study is complicated by various imperfections in the data, which we emphasize below. We hope to contribute to better annotation of RMs in future releases of UD; in particular, we have the following goals:

- to present an overview of syntactic and semantic functions of RMs in Slavic languages and to examine their current and desired annotation;
- to present a brief review of selected RMs in Romance and Germanic languages;
- to propose improvements in order to make the data more consistent;
- possibly also to suggest how the guidelines could be made more transparent.

We do not discuss the appropriateness of classifying the RM as a pronoun (in languages where it has a form of a reflexive clitic). In our opinion, this is a controversial question, but we are not sure that a more widely acceptable solution exists.

2 Detecting Reflexives in the UD Treebank Data

In order to make any cross-linguistic claims about reflexives, one must be able to recognize them in corpora. In Universal Dependencies, reflexive words should be annotated with the feature Reflex=Yes. Hence, the feature is our primary source of information; but it apparently has not been used everywhere it should. In UD 2.2, 56 treebanks use the feature.[1] Out of the 71 languages covered in UD 2.2, the Reflex

[1] http://universaldependencies.org/ext-feat-index.html#reflex

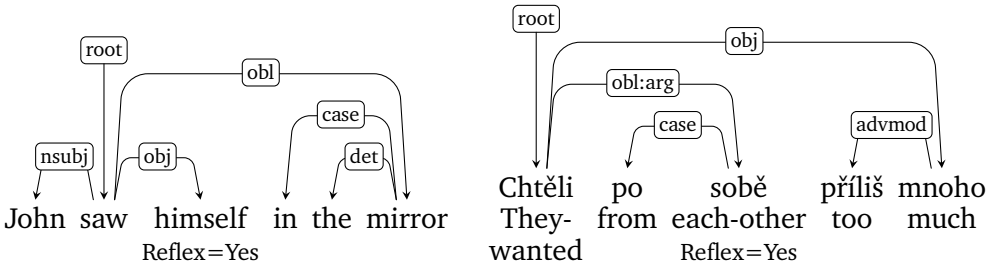

Figure 3: True reflexive (left) and reciprocally used reflexive (right).

feature is present in 44 languages from various language families (Indo-European, Afro-Asiatic, Uralic, Turkic, Mongolic, Dravidian).

In the treebanks where Reflex=Yes is found, it mostly occurs with the PRON part of speech (pronoun) or with DET (determiner; used for reflexive possessives in some languages). A few treebanks use the feature with verbs, adjectives or particles.

There are languages that have reflexive pronouns, yet their treebanks fail to mark them with Reflex=Yes. The list omits, for instance, Hindi, although Hindi has reflexive pronouns such as *apne*. Our observations in the present work are mostly limited to treebanks where the feature is used.

3 Relations / Constructions

3.1 Core and Oblique Dependents

In the simplest case, a reflexive pronoun is used as an object or an oblique dependent of the verb. An irreflexive personal pronoun could occur in the same position, and the syntactic structure is the same regardless of whether the pronoun is or is not reflexive. In the present work, we refer to these cases as *true* or *semantic reflexivity*.

In some languages reflexive pronouns are used also with reciprocal meaning (other languages have dedicated reciprocal pronouns such as "each other" in English). For instance in German

(2) *Braut und Bräutigam haben* ***sich*** *geküsst.* "The bride and groom kissed **each other**."

Here the reflexive pronoun is coreferential with the plural subject as a whole. With regard to semantic roles, each of the two individuals takes the role of the subject and the object at the same time; syntactically however, they function as the subject and the RM as object or oblique dependent (see the second example in Figure 3).

The UD annotation guidelines do not distinguish true reflexives from reciprocally used RMs. They distinguish whether the RM occurs in place of a direct object (labeled

obj), indirect object (iobj) or oblique dependent (labeled obl or one of its subtypes such as obl:arg).

3.2 Inherently Reflexive Verbs

Certain verbs in certain languages require a reflexive clitic without assigning it any semantic role or syntactic function. Traditional grammar treats the clitic as a free lexical morpheme, which is part of the verbal lexeme.

For example, the Slovenian verb *smejati se* "to laugh" is inherently reflexive and never occurs without *se* (see Figure 2). Here the UD guidelines specify that the RM shall be attached as expl:pv (standing for "expletive, subtype pronominal verb").[2]

Some authors restrict the class of inherently reflexive verbs to those that do not have irreflexive counterparts from which they can be derived (Svoboda, 2014, p. 11). Others claim that many transitive verbs have an inherently reflexive variant that is not semantically equivalent to the transitive form, though it is related to it in some way. However, the nature of this semantic relation cannot always be captured easily (Waltereit, 2000, p. 257). Hence it may be difficult to evaluate whether its meaning is "different enough" to rule out the possibility that the reflexive pronoun is just a normal object.

Sometimes a translation into another language may provide a clue. For example, the Latvian *mācīt kam* corresponds to English "to teach somebody", while the reflexive *mācīties* corresponds to "to learn". The difference in translation suggests that the reflexive verb is semantically different and can thus be considered a derived inherently reflexive verb. A more general (but vaguer) clue results from comparing the action performed by the actor of the verb: in Bulgarian *върна нещо някъде (vărna nešto njakăde)* "to return something somewhere", the actor takes an object and moves it in space; even if the object is animate and can move independently, some physical or mental force is applied to make it move in the right direction, quite possibly against its own will. On the other hand, *върна се някъде (vărna se njakăde)* "to return somewhere" describes independent and free movement of the actor to a place where they have been in the past. Again, we can conclude that it describes a different action and is no longer transitive.

The decision is easier if the irreflexive counterpart is intransitive, as in Spanish *ir* "to go" (irreflexive, intransitive) vs. *irse* "to leave" (inherently reflexive).

3.3 Reflexive Passives and Impersonal Constructions

Reflexives in some languages have grammaticalized into markers of alternations in voice (diathesis). They serve as additional means of expressing passive of transitive verbs (especially if their object is inanimate): the original object becomes subject, the

[2] http://universaldependencies.org/u/dep/expl.html#Reflexives

original subject is removed, and a reflexive appears in object position. If a reflexive clause of the form "X did itself" can be paraphrased as "X was done" or "Somebody did X", it is a reflexive passive.[3] The guidelines specify that the RM be attached by the relation `expl:pass` in these cases. For example, in Upper Sorbian

(3) *Serbski institut je so k 1. januarej 1992 wot Swobodneho stata Sakska wutworił.* "The Sorbian Institute was created on January 1, 1992 by the Free State of Saxony."

the RM *so* forms a reflexive passive.

Although the construction is presented as a variant of the passive voice, the verb is actually in its active form; as (Sussex and Cubberley, 2011, p. 448) note, the verb-reflexive complex "is not unlike the Greek middle voice in function and meaning."

Another category, representing another shade of meaning, is called in the literature anticausative (Svoboda, 2014, p. 1–2), decausative or inchoative (Silveira, 2016, p. 116). Examples include Czech and Portuguese

(4) *Dveře se otevřely.* "The door opened."

(5) *O vaso se quebrou.* "The vase broke."

There must be an external cause of the event but the cause is unknown or unidentifiable, and the event seems to come about spontaneously. The difference between anticausative, middle voice and reflexive passive is semantic rather than syntactic, very subtle and hard to discern. Hence it does not seem to be something that can or should be distinguished in UD; we will call them all 'reflexive passive'.

Reflexives can also be used in impersonal constructions, i.e., clauses without subject. They resemble very much the reflexive passive except that the verb is not transitive and there is no object that could be promoted to the subject position. Consequently, the default agreement is triggered on the verb. What exactly it means is language-dependent; for example, in Slavic languages it is the third person, singular, neuter—cf. Polish

(6) *Po Edenie chodziło się nago.* (lit. *Along Eden walked itself nude.*) "One would walk nude in Eden." (Patejuk and Przepiórkowski, 2015).

[3] While the term reflexive passive is used in some classical grammars, other authors regard it as controversial, pointing out differences between reflexive and periphrastic passivization. The stance depends on what one considers the defining properties of passive. In UD, the primary purpose of the `:pass` relation subtype is to signal non-canonical mapping between syntactic relations and semantic roles: auxiliaries in periphrastic passives use `aux:pass`, reflexives use `expl:pass`.

	Sing	Plur
1(|Reflex)	*mir, mich*	*uns*
2(|Reflex)	*dir, dich*	*euch*
3|Masc	*ihm, ihn*	*ihnen, sie*
3|Fem	*ihr, sie*	*ihnen, sie*
3|Neut	*ihm, es*	*ihnen, sie*
3|Reflex	*sich*	*sich*

Table 1: German object and reflexive pronouns. The two forms are dative and accusative; some pronouns have one form for both cases.

The reflexive marker should be attached as `expl:impers` in these cases.

Silveira (2016, p. 124) notes that in some languages impersonal construction can contain even a transitive verb. The 'internal argument' (object) then does not become subject (unlike in passive). It keeps the accusative case and does not trigger verb agreement in person and number; the verb stays in the default third person singular: Spanish

(7) *Se observa cambios en la economia.* "Changes are observed in the economy."

3.4 Double Function / Haplology

A reflexive marker can have a double function. For instance, if an inherently reflexive verb is used in an impersonal construction, there will be just one RM, not two (Patejuk and Przepiórkowski, 2015). Another point is that sometimes one reflexive is shared by two verbs, as in Slovenian, with one inherently reflexive verb controlling another:

(8) *Bal se je smejati.* "He was afraid to laugh."

The guidelines currently do not specify the preferred solution of such cases.

4 Reflexives in Germanic Languages

While English has two parallel sets of irreflexive and reflexive pronouns, in many other languages only the third person has a special reflexive form. First and second person pronouns have the same form whether they are used reflexively or not, but the reflexive usage can be easily recognized because the verb form and the subject pronoun is in the same person and number as the object pronoun. Table 1 shows the pronouns in German.

Table 2 shows that 13 out of 19 Germanic treebanks use the Reflex feature. The remaining 6 treebanks indeed contain reflexive pronouns but they do not tag them as such. Only 6 treebanks use the feature also with 1st and 2nd person pronouns,

Treebank	R12	R3	obj	iobj	obl	expl:pv
Afrikaans	4	3	41		14	
Danish		34	77	12	8	
Dutch Alpino		26	22	2	5	69
Dutch LassySmall		22	28	0	3	69
English EWT	3	2	52	8	15	
English GUM	3	3	54	10	23	
English LinES	4	17	49	2	23	
English PUD		5	80		20	
Faroese		18	67		28	
German GSD	3	55	73	8	3	12
Gothic	1	37	18	9	36	
Norwegian Bokmaal		40	69	14	13	
Norwegian NynorskLIA		16	77	5	14	

Table 2: **R12** = Number of first and second person non-possessive reflexives per 10,000 tokens. **R3** = Number of third person non-possessive reflexives per 10,000 tokens. **obj** = Percentage of reflexives attached as objects. **iobj** = % indirect objects. **obl** = % oblique arguments and adjuncts. **expl:pv** = % inherently reflexive verbs.

although these pronouns are sometimes used reflexively in the other treebanks, too.

Most RMs are attached as true reflexive arguments (`obj`, `iobj`, `obl`). Only in German and Dutch we see some inherently reflexive verbs. They should probably appear elsewhere too, with the exception of English. For instance, we believe that Norwegian *Jeg føler meg som…* "I feel (myself) like…" should be inherently reflexive.

5 Reflexives in Romance Languages

The situation is even less satisfactory in the Romance languages (Table 3). Only 7 treebanks out of 21 tag their RMs with Reflex=Yes. In addition, the Italian ISDT treebank does not use the feature but it uses the reflexive-specific relations `expl:pass` and `expl:impers`. Furthermore, two French treebanks use Reflex=Yes for disjoint sets of pronouns: Sequoia for the clitics *se, me, te, nous, vous,* GSD for the tonic reflexives such as *lui-même, eux-mêmes*.

Some treebanks label all reflexive clitics as core arguments: Spanish GSD as indirect objects, Italian ParTUT (and according to Silveira (2016, p. 126, 129) also French GSD) as direct objects. It is obvious that a significant number of instances are non-argumental and should thus be labeled as expletives.

The Sequoia treebank of French uses just the universal label `expl` for all non-argumental RMs and does not further distinguish inherent, passive and impersonal constructions. This solution is advocated by Silveira (2016, p. 143) who says that the distinction is a source of uncertainty for both parsers and annotators.

Treebank	R12	R3	obj	iobj	obl	expl	:pv	:pass	:imp
French GSD	0	2	2		50				
French Sequoia	5	43	9	3		88			
Italian PUD		1			67				
Italian ParTUT		3	100						
Romanian Nonstd		30	0	2	2		71	15	4
Romanian RRT		180	2	2	0	0	53	28	3
Spanish GSD	2	123	0	99	0				

Table 3: Romance treebanks. **R12**, **R3**, **obj**, **iobj**, **obl** and **expl:pv** – see Table 2. **:pass** = % passive RMs. **:imp** = % impersonal RMs.

Romanian differs from the other Romance languages by having two sets of reflexive pronouns, one in the accusative, and one in the dative (Cojocaru, 2003). The Romanian treebanks do distinguish various subtypes of `expl` as defined by the guidelines and summarized in Sections 3.2 and 3.3. Reflexives analyzed as core arguments are much less frequent than expletives. There is also a Romanian-specific relation, `expl:poss`, used for possessive or benefactive meaning of dative clitics (note that this construction is different from the reflexive possessives mentioned in Section 1).

6 Reflexives in Slavic Languages

Slavic personal pronouns, including reflexive pronouns, have full and clitic forms. Unlike Germanic and Romance languages, the same form is used in all persons and numbers. The full forms occur in all cases except nominative and vocative, and are used to encode true reflexivity, i.e., a nominal that is coreferential with the subject of the clause. The clitic forms are used as true reflexives, too, but they often have other grammatical functions listed in Section 3. There are one or two reflexive clitics per language, which can be characterized as accusative *(se/sa/so/się/sę)* and dative *(si/sej)* forms. In East Slavic (and also in Baltic) languages the reflexive clitic has become a suffix of the verb (Geniušienė, 1987, p. 241).

Clitics and full pronouns are semantically equivalent when they function as true reflexives, and the choice of one of them over the other is made for prosodic and/or pragmatic reasons (topic-focus articulation). The default form for expressing true reflexivity in South and West Slavic languages is the clitic form, while the full form is reserved for expressing emphasis or contrast, when the reflexive pronoun is coordinated with another noun phrase, or after prepositions. The full form can only be used in truly reflexive and reciprocal contexts (Svoboda, 2014, p. 5–6).

6.1 Annotation in Slavic Treebanks

Table 4 gives a summary of RMs in West and South Slavic treebanks. Fourteen treebanks (out of 15) use the Reflex=Yes feature. The first observation is that RMs

Treebank	RM	obj	iobj	obl	expl	:pv	:pass	:imp
Bulgarian	205	0	0	0	98			
Croatian	146	16	0	0		75		
Czech CAC	183	1	1	7		67	23	
Czech CLTT	133	0		1		24	74	
Czech FicTree	366	5	3	10		75	6	
Czech PDT	171	5	2	6		67	19	
Czech PUD	190	9	2	5		68	13	
Old Church Slavonic	28	13	8	77				
Polish LFG	272	2	4	4		85		4
Polish SZ	225	2	2	4		91		
Serbian	140	0		1	98			
Slovak	290	1	1	4		85	7	
Slovenian SSJ	172	1	0	3	95			
Upper Sorbian	178	8		1		40	50	

Table 4: Slavic treebanks. **RM** = Number of non-possessive reflexives per 10,000 tokens. Other columns as in Table 2; `compound` instead of `expl` in Serbian.

are much more frequent than in the Germanic and Romance languages. The counts include both clitics and full pronouns, but the latter are only a small fraction of the whole. Two thirds or more are the occurrences of the clitic *se*.

We do not include East Slavic treebanks in the overview because reflexive clitics are not independent words there. The current UD guidelines actually assume that reflexive verbs should be treated as multi-word tokens, consisting of the verb proper and the clitic suffix. For example, Russian *проснуться (prosnut'sja)* "to wake up" would be split to two syntactic words, *prosnut'* and *sja*, which would make it parallel to the other Slavic languages (e.g. Czech *vzbudit se* "to wake up"). However, this is not what we find in the data. The reflexive verbs are kept together and marked either with Voice=Mid (in Russian and Belarusian) or nothing at all (in Ukrainian). Note that the corresponding verb in Spanish is also sometimes[4] written together with the clitic (*despertarse* "to wake up") but in UD it is split into two words *(despertar se)*. East Slavic languages express true reflexivity using the full pronoun, *sebja* (Svoboda, 2014, p. 29). However, although *-sja* no longer functions as a truly reflexive marker, it can mark the other functions which we investigate in this study.

The next observation is that Old Church Slavonic (OCS) is an outlier. Only true reflexives are tagged as pronouns with Reflex=Yes. The vast majority of occurrences of *sę* are tagged `AUX` and attached via the dependency relation `aux`. This is perhaps inherited from the original PROIEL annotation, but it does not seem to be in accord with the UD guidelines. Supposedly, these occurrences of *sę* function as the RMs for

[4]In the infinitive and imperative.

inherent reflexives, passives and impersonal constructions.

As for dependency relations, the non-AUX RMs in OCS are attached as obj, iobj or obl, which supports the hypothesis that only true reflexives are annotated in the way described in Section 3.

Three other languages (Bulgarian, Serbian and Slovenian) make no difference between the various functions of the RM *se*. They uniformly attach it as expl (in case of Bulgarian and Slovenian) or compound (Serbian). Samardžić et al. (2017, p. 41, 42) explain the reasoning behind their use of compound for all instances of *se*. They view this form as a detachable morpheme belonging to the verb to which it is attached both in lexical and morphological sense. In their view, *se* is not just a prosodic variant of the full reflexive pronoun. In fact, they claim that it is not a pronoun at all, and consequently, it should be analyzed in the same way in all its uses: as a free morpheme marking absence of one of the verb's core arguments. They also state that the different functions noticeable in the other treebanks are higher-level interpretations of the same syntactic form, which should not be part of UD. While we can agree that in many cases this is true, there are still cases where *se* is a true reflexive pronoun, that is, both full and clitic form are possible and it is commutable with clitics of irreflexive personal pronouns. Furthermore, if a single relation is used for all functions other than true reflexives and reciprocals, expl seems to be a better solution than compound, as argued by Silveira (2016) and codified by the UD guidelines.

Figures 4 and 5 show the transitive verb *smatrati* "to consider" in Serbian and Croatian, in both cases in a passive construction. The "compound everywhere" approach is not able to distinguish this from an inherently reflexive verb. Figure 6 is an example of *se* functioning as a core object. We believe that the corresponding Serbian sentence should use obj here too (instead of compound). Therefore, the annotation should differentiate between true reflexivity and the other functions *se* can have.

Polish distinguishes expl:impers but not expl:pass. On the other hand, Czech, Slovak, Upper Sorbian and Croatian distinguish the passive RMs (Croatian labels them aux:pass instead of expl:pass), but they do not use expl:impers. We would argue that both types of constructions exist in all these languages. For example, the Czech FicTree treebank contains impersonal *zapomnělo se na ně* "they were forgotten". In Polish LFG, the anticausative *Drzwi zamknęły się* "The door closed" could be analyzed as expl:pass but has expl:pv instead. However, genuinely passive examples in Polish seem to be rare, presumably because Polish favors the impersonal constructions that keep objects in the accusative: *Maluje się ściany* "The walls are painted."

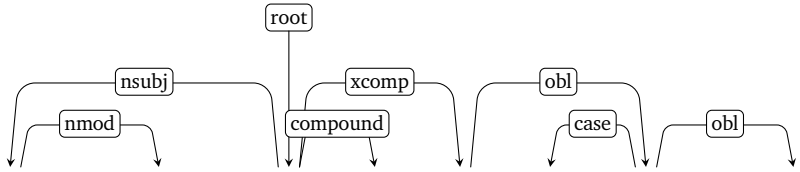

Figure 4: The Serbian treebank always attaches *se* as `compound`. Here, its real function is the reflexive passive.

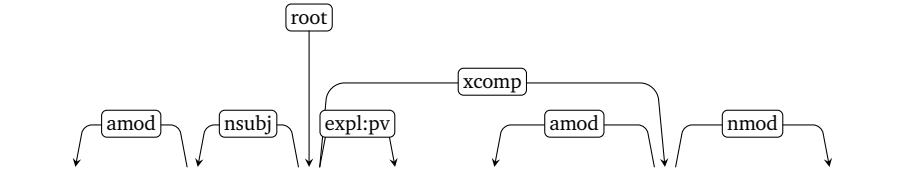

Figure 5: In the Croatian treebank, this *se* is annotated with `expl:pv`, although arguably it has the passive meaning.

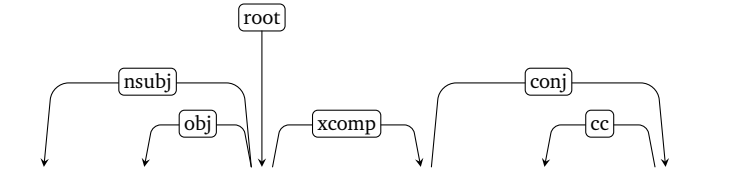

Figure 6: This example is from Croatian but its Serbian counterpart would be similar. Here, *se* is commutable with the full reflexive *sebe*. Hence *se* should be `obj` in this context, in both languages. (We had to shorten the tree due to lack of space, but the full sentence rules out potential ambiguity with `expl:pass`, as *osoba* is the actor: *Osoba ne nalazi nikakvo uporište u sebi samoj, ne vjeruje u sebe, dapače, smatra se bezvrijednom i beznačajnom.* “The person does not find any strength in herself, does not believe in herself, in fact, considers herself worthless and insignificant.”)

7 How to Improve the Treebanks

Annotating all RMs with Reflex=Yes should be easy to achieve because the feature is often tied to just a few lemmas (but it is still helpful for users who do not know the lemmas). Germanic and Romance languages should consider disambiguating reflexive usages of 1st and 2nd person pronouns, as it is already done in German.

As a minimum, all treebanks should distinguish between true reflexivity (`obj` / `iobj` / `obl`) and the non-argumental RMs (`expl` and subtypes). The distinction can often be based on lists of verb lemmas, but it still means that significant manual effort is needed in Bulgarian, Serbian and Slovenian, as well as in some Germanic and Romance languages.

Distinguishing various subtypes of `expl` is optional in UD, yet we would like to advocate at least the distinction between `expl:pv` as a lexical morpheme on one side, and `expl:impers` or `expl:pass` as grammatical means on the other side. Clearer instructions for identifying inherently reflexive verbs are needed, and we have proposed some heuristics in Section 3. The UD guidelines must be modified if East Slavic languages shall keep their reflexive verbs as single syntactic words. The middle voice should then be used in all three languages (it is currently not used in Ukrainian) in order to live up to the UD motto that "same thing should be annotated same way."

Croatian `aux:pass` should be replaced by `expl:pass` and the current `aux` instances should be fixed manually. In Old Church Slavonic, the `AUX/aux` annotation should be replaced by labels that follow the guidelines.

The UD guidelines should specify the priorities when *se* has a double function or is shared by two verbs. In general, the guidelines should provide a broader overview of reflexives with examples from multiple languages; at present, the relevant rules are scattered under various labels.

8 Conclusion

We have shown that the annotation of reflexives in Universal Dependencies is currently unsatisfactory, inconsistent and unpredictable. There is a range of possible causes. First and foremost, constructions with reflexives are a difficult and sometimes controversial issue in many European languages. Multiple analogies with other parts of grammar are available, but most of them have their downsides too. Maintainers of UD treebanks often disagree in their choice of annotation options. More attention should be paid to reflexives in the guidelines and there should be a section devoted to reflexives and discussing all their functions, with examples from many languages. Finally, the data providers should be encouraged to follow a single interpretation of the guidelines, especially in cases where the current annotation can be fixed by an automated procedure.

Acknowledgments

We would like to thank the anonymous reviewers for useful comments. The research was partially supported by the SVV project number 260 453, the grant 15-10472S of the Czech Science Foundation (GAČR), and FP7-ICT-2009-4-249119 (MŠMT 7E11040).

References

Dana Cojocaru. 2003. *Romanian Grammar*. Slavic and East European Language Research Center (SEELRC), Duke University.

Emma Geniušienė. 1987. *The typology of reflexives*. Mouton de Gruyter, Berlin, Germany.

Václava Kettnerová and Markéta Lopatková. 2014. Reflexive verbs in a valency lexicon: The case of Czech reflexive morphemes. In *Proceedings of the XVI EURALEX International Congress: The User in Focus*, Bolzano/Bozen, Italy.

Ekkehard König and Peter Siemund. 2000. Intensifiers and reflexives. a typological perspective. In Zygmunt Frajzyngier and Traci S. Curl, editors, *Reflexives: Forms and Functions. Volume 1*. John Benjamins Publishing Company, Amsterdam/Philadelphia.

Joakim Nivre, Mitchell Abrams, Željko Agić, Lars Ahrenberg, Lene Antonsen, Maria Jesus Aranzabe, Gashaw Arutie, Masayuki Asahara, Luma Ateyah, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Elena Badmaeva, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, John Bauer, Sandra Bellato, Kepa Bengoetxea, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Rogier Blokland, Victoria Bobicev, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Aljoscha Burchardt, Marie Candito, Bernard Caron, Gauthier Caron, Gülşen Cebiroğlu Eryiğit, Giuseppe G. A. Celano, Savas Cetin, Fabricio Chalub, Jinho Choi, Yongseok Cho, Jayeol Chun, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Marine Courtin, Elizabeth Davidson, Marie-Catherine de Marneffe, Valeria de Paiva, Arantza Diaz de Ilarraza, Carly Dickerson, Peter Dirix, Kaja Dobrovoljc, Timothy Dozat, Kira Droganova, Puneet Dwivedi, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Tomaž Erjavec, Aline Etienne, Richárd Farkas, Hector Fernandez Alcalde, Jennifer Foster, Cláudia Freitas, Katarína Gajdošová, Daniel Galbraith, Marcos Garcia, Moa Gärdenfors, Kim Gerdes, Filip Ginter, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta Gonzáles Saavedra, Matias Grioni, Normunds Grūzītis, Bruno Guillaume, Céline Guillot-Barbance, Nizar Habash, Jan Hajič, Jan Hajič jr., Linh Hà Mỹ, Na-Rae Han, Kim Harris, Dag Haug, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Jena Hwang, Radu Ion, Elena Irimia, Tomáš Jelínek, Anders Johannsen, Fredrik Jørgensen, Hüner Kaşıkara, Sylvain

Kahane, Hiroshi Kanayama, Jenna Kanerva, Tolga Kayadelen, Václava Kettnerová, Jesse Kirchner, Natalia Kotsyba, Simon Krek, Sookyoung Kwak, Veronika Laippala, Lorenzo Lambertino, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phuong Lê Hồng, Alessandro Lenci, Saran Lertpradit, Herman Leung, Cheuk Ying Li, Josie Li, Keying Li, KyungTae Lim, Nikola Ljubešić, Olga Loginova, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Héctor Martínez Alonso, André Martins, Jan Mašek, Yuji Matsumoto, Ryan McDonald, Gustavo Mendonça, Niko Miekka, Anna Missilä, Cătălin Mititelu, Yusuke Miyao, Simonetta Montemagni, Amir More, Laura Moreno Romero, Shinsuke Mori, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Luong Nguyễn Thị, Huyền Nguyễn Thị Minh, Vitaly Nikolaev, Rattima Nitisaroj, Hanna Nurmi, Stina Ojala, Adéday`ọ Olúòkun, Mai Omura, Petya Osenova, Robert Östling, Lilja Øvrelid, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Siyao Peng, Cenel-Augusto Perez, Guy Perrier, Slav Petrov, Jussi Piitulainen, Emily Pitler, Barbara Plank, Thierry Poibeau, Martin Popel, Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Tiina Puolakainen, Sampo Pyysalo, Andriela Rääbis, Alexandre Rademaker, Loganathan Ramasamy, Taraka Rama, Carlos Ramisch, Vinit Ravishankar, Livy Real, Siva Reddy, Georg Rehm, Michael Rießler, Larissa Rinaldi, Laura Rituma, Luisa Rocha, Mykhailo Romanenko, Rudolf Rosa, Davide Rovati, Valentin Roșca, Olga Rudina, Shoval Sadde, Shadi Saleh, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Baiba Saulīte, Yanin Sawanakunanon, Nathan Schneider, Sebastian Schuster, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Mo Shen, Atsuko Shimada, Muh Shohibussirri, Dmitry Sichinava, Natalia Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Kiril Simov, Aaron Smith, Isabela Soares-Bastos, Antonio Stella, Milan Straka, Jana Strnadová, Alane Suhr, Umut Sulubacak, Zsolt Szántó, Dima Taji, Yuta Takahashi, Takaaki Tanaka, Isabelle Tellier, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Francis Tyers, Sumire Uematsu, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Sowmya Vajjala, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Veronika Vincze, Lars Wallin, Jonathan North Washington, Seyi Williams, Mats Wirén, Tsegay Woldemariam, Tak-sum Wong, Chunxiao Yan, Marat M. Yavrumyan, Zhuoran Yu, Zdeněk Žabokrtský, Amir Zeldes, Daniel Zeman, Manying Zhang, and Hanzhi Zhu. 2018. Universal dependencies 2.2. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the 10th International Conference*

on Language Resources and Evaluation (LREC 2016), pages 1659–1666, Portorož, Slovenia. European Language Resources Association.

Agnieszka Patejuk and Adam Przepiórkowski. 2015. An LFG analysis of the so-called reflexive marker in Polish. In *The Proceedings of the LFG'15 Conference*, pages 270–288, Stanford, CA, USA. CSLI Publications.

Tanja Samardžić, Mirjana Starović, Željko Agić, and Nikola Ljubešić. 2017. Universal Dependencies for Serbian in comparison with Croatian and other Slavic languages. In *Proceedings of the 6th Workshop on Balto-Slavic Natural Language Processing*, pages 39–44, Valencia, Spain. Association for Computational Linguistics.

Natalia G. Silveira. 2016. *Designing syntactic representations for NLP: An empirical investigation*. Stanford University, Stanford, CA, USA.

Roland Sussex and Paul Cubberley. 2011. *The Slavic Languages*. Cambridge Language Surveys. Cambridge University Press.

Ondřej Svoboda. 2014. *Functions of the Czech reflexive marker se/si (research master's thesis)*. Universiteit Leiden, Leiden, Netherlands.

Richard Waltereit. 2000. What it means to deceive yourself: The semantic relation of French reflexive verbs and their corresponding transitive verbs. In Zygmunt Frajzyngier and Traci S. Curl, editors, *Reflexives: Forms and Functions. Volume 1*. John Benjamins Publishing Company, Amsterdam/Philadelphia.