To appear in
Russian Linguistics (2025)

Björn Wiemer (wiemerb@uni-mainz.de), Joanna Wrzesień-Kwiatkowska (kwiatkow@uni-mainz.de), Marek Łaziński (m.lazinski@uw.edu.pl), Alexander Rostovtsev-Popiel (rosto@uni-mainz.de), Rafał L. Górski (rafal.gorski@ijp.pan.pl),
Katarzyna Osior-Szot (kos@fundacjajezykapolskiego.pl)

Database of aspect triplets in Polish, Czech and Russian – Presentation, considerations and case studies –

Supplement to subsection

3.2.2. Problems caused by reflexive markers

In connection with the assessment of token frequencies a serious problem arises with the treatment of the reflexive marker. Our database contains many units that are distinguished in form by the presence vs absence of a reflexive marker, and this corresponds to a difference in lexical meaning. For instance, in the Polish part of the database we have both *barwić* 'color' and *barwić się* 'take on color' (each in two triplets), but we only have transitive *cechować* 'characterize'; the lexical unit *cechować się* 'be characterizable' (with reflexive marker) does exist, but this anticausative derivative of *cechować* did not make it into the database.[1] Simultaneously, the reflexive marker can have very different functions, among others a grammatical function which may apply to almost any transitive verb (like *cechować*). Consequently, one would have to check for every corpus token whether the reflexive marker "belongs" to a transitive verb (*barwić*, *cechować*, etc.) or to its intransitive ("reflexive-marked") derivative (*barwić się*, *cechować się*, etc.).

In Polish (*się*) and Czech (*se*) this marker is an enclitic, while in Russian it is a postfix, i.e. a bound morpheme which takes the last position in a word form (*-sja*, with an allomorph *-s'*). In Kemmer's (1993) terms, all these markers count as light reflexive markers. They raise problems connected to the morphology—lexicon interface, and they have severe consequences when it comes to inventorizing lexical units (with/without the reflexive marker), but even much more so for any attempt at determining token-frequencies of these units from corpora.

The problem: lack of discriminating annotation

In Polish and Czech, the reflexive enclitic is employed in the derivation of lexical units, such as anticausatives (Pol. *tworzyć* 'create' ⇒ *tworzyć się* 'arise', *wyczerpać* 'exhaust' ⇒ *wyczerpać się* 'get exhausted'), autocausatives (Pol. *brudzić* 'stain, make dirty' ⇒ *brudzić się* 'get, make oneself dirty', *zobowiązać* 'oblige (sb)' ⇒ *zobowiązać się* 'oblige oneself'), deaccusatives (*podjąć (pracę*.ACC*)* 'start (work)' ⇒ *podjąć się (zadania*.GEN*)* 'take on (a task)') and a few other smaller groups.[2] In these cases, the reflexive enclitic changes the argument structure of the corresponding non-reflexive verb and can thus be regarded as a component of a distinct lexical unit. For this reason, these units are given separate entries in dictionaries.[3] In addition, the reflexive enclitic abundantly functions as a marker of voice operations that do not change argument structure, but only reorganize the morphosyntactic coding of arguments. The

[1] Obviously, because it only has a stative meaning and thus no pfv. partner (from which a IPFV2 might be derived).
[2] Cf. Wiemer 2007 for Polish, also Holvoet (2020) for an overview of these groups and their relation to reflexive and reciprocal meanings in Baltic and Slavic.
[3] In semantic and diachronic terms, derivation not always starts from the transitive stem, and there are a couple of *reflexiva tantum* (e.g., Pol. *bać się* 'be afraid'). However, for the concerns of the database this is irrelevant.

latter operations do not create new lexical units. First of all, this applies to so-called impersonal constructions (a.k.a. backgrounding passive), in which the most agent-like argument is syntactically demoted, but no further changes in argument coding (let alone in argument structure) are involved.[4] Consider, for instance, the most prominent case in Polish (often referred to as 'impersonal *się*', Pol. SIĘ *bezosobowe*):

(1a) *Wszysc-y* ***myj-ą*** *zęb-y* *po* *kolacj-i*.
everybody-NOM.PL:VIR wash[IPFV]-PRS.3PL tooth-ACC.PL after dinner-LOC
'Everybody **cleans** their teeth after dinner.'

(1b) *Zęb-y* ***myj-e*** ***się*** *po* *kolacj-i*.
tooth-ACC.PL wash[IPFV]-PRS.3SG RM after dinner-LOC
'**One cleans** one's teeth after dinner.' (≅ 'Teeth **are cleaned** after dinner.')

The same type is well-known in Czech as well; see (2):

(2) *Sumíčkovití zahrnují asi 210 druhů ve 30 rodech, a jsou tak velmi rozsáhlou skupinou, jejíž systematické postavení rozhodně není definitivní.*
Tradičně ***se*** *čeleď* ***rozděluj-e*** *na* *pět* *podčeled-í*.
traditionally RM family-(NOM.SG) divide[IPFV]-PRS.3SG on five subfamily-GEN.PL
'Sumic-likes include about 210 species in 30 genera, and are thus a very large group whose systematic position is decidedly not definitive. Traditionally the family **is divided** into five subfamilies.'

Moreover, the reflexive marker is used in true reflexives and reciprocals (both natural and canonical ones).[5] These also do not change the number of arguments, but only restructure their relation to the involved referents and their coding in the syntax; for instance,

(3) Pol. *Waldek za wszystko* ***obwiniał się*** *sam*. reflexive proper
'For everything Waldek **accused himself**.'

(4) Pol. *Liderzy partii* ***obwiniali się*** *za niedociągnięcia*. canonical reciprocal[6]
'The party leaders **blamed each other** for the shortcomings.'

(5) Cz. *Manželé* ***se objímali***. natural reciprocal
'The spouses **embraced (each other)**.'

As a shortcut, we may distinguish these two functions of the reflexive enclitic by dubbing them 'lexical *się/se*' (> used to derive lexical units) and 'grammatical *się/se*' (> used in voice-related functions). Strictly speaking, the latter includes true reflexives and reciprocals; however, many dictionaries treat them as separate lexical entries (a reason seems to be their frequent occurrence, particularly in coding reciprocal situations).[7] In such cases, we abided by dictionary practice and introduced *se/się*-units as separate entries.

For Russian, similar problems with the reflexive marker arise to a lesser extent. As a postfix, *-sja*/*-s'* is bound to word forms, in the RNC (and other corpora) one can therefore search for them separately from word forms and lemmata without this affix. Nonetheless, there remains a

[4] Cf. Rytel-Kuc (1990) and Górski (2008: 50-55, 82f.) for overviews. These constructions consistently imply that the demoted argument denotes a human.

[5] For this distinction in Polish cf. Wiemer (1999; 2007), following Kemmer (1993) and Nedjalkov (2007: 7-10), who treats canonical reciprocals as "standard reciprocal opposition".

[6] In Polish and Czech (as in German), many canonical reciprocals systematically alternate with true reflexive readings. Thus, under favorable circumstances, (4) may also be understood as '…blamed themselves (collectively)'. For some discussion concerning Czech cf. Panevová et al. (2020).

[7] For an account of the research history behind these distinctions and their relationship to voice alternations cf. Wiemer (forthc.).

problem in distinguishing transitive verbs from their reflexive-marked counterparts, since *sja*-marked forms are employed in passive constructions (see ex. 18 in the main article) which, for ipfv. stems, are often "homonymous" with anticausatives (compare 6-7) or autocausatives (compare 8-9).

(6) *Zarubežnymi učenymi ėtot vopros **podnimalsja** ešče v načale 1990-x godov, rossijskie tol'ko pristupajut k interpretacii ėtogo javlenija*.
'This issue **was raised** by foreign scientists in the early 1990s, while Russian scientists are just beginning to interpet this phenomenon.' → passive of transitive *podnimat'* 'raise, lift'
(RNC; *Informacionnoe obščestvo*. 2014)

(7) *Ėtot šum **podnimalsja** snizu k nogam i v lico prokuratoru*.
'This noise **rose** from below to the feet and in the face of the procurator.' → anticausative
(RNC; M.A. Bulgakov: *Master i Margarita*. 1929-1940)

(8) *Otnositel'no nizkaja cena byla obuslovlena tem, čto **sobiralsja** Ford Focus takže na rossijskix zavodax*.
'The relatively low price was due to the fact that the Ford Focus was also **put together** at Russian factories.' → passive of transitive *sobirat'* 'put together, assemble'
(RNC; *Zerkalo mira*. 2012)

(9) *V trudnye gody antialkogol'nogo makkartizma, v podpol'e, ljudi **sobiralis'** gruppami po troe, čtob posmotret' ėtot fil'm ili daže prosto ponjuxat' kassetu...*
'In the difficult years of anti-alcohol McCarthyism, in the underground, people **gathered** in groups of three to watch this film or even just sniff the cassette...' → autocausative
(RNC; G. Gorin: *Ironičeskie memuary*. 1990-1998]

Again, corpora do not tag the function of the reflexive marker, as it can hardly be done automatically. As a consequence, functions in the derivation of separate lexemes (certainly more frequent) and voice-like functions (probably rarer) cannot be told apart (as for the RNC, cf. Goto/Say 2009).

This situation poses a big problem for assessing token frequencies. Our database contains many "pairs" of a transitive stem with its reflexive-marked counterpart, but the different functions of the reflexive marker are not annotated in the PNC, CzNC or RNC (nor in any other corpus we know of). While the database entries (= types = potential aspect triplets) were compiled from a systematic survey of dictionaries and similar sources (see §3.1), on the "surface" of text strings these two units cannot be distinguished and suitable tags are lacking. It is therefore impossible to obtain reliable token frequencies of such pairs included in the database. This would require manual annotation, which is unfeasible for any larger samples. To our knowledge, so far nobody has managed to solve this problem, nor have we been able to do so. In addition, as for Polish and Czech, though in most cases the enclitic is positioned close to the verb, there are occurrences where they can be quite remote from each other. Therefore, verb forms with *się/se* require a retrieval with sophisticated queries.

Assessing the significance of this problem in the triplet database

Being unable to circumvent the problem, we at least tried to assess its significance for the entries in the database. The procedure was complicated, and we could perform it only for the Polish part of the database (due to lack of personal resources). Here we report on its general outline. For all ipfv. stems in the triplet database that are attested with and without *się* as separate

lexemes (e.g., *tworzyć* 'create' – *tworzyć się* 'arise')[8] we ran corpus queries[9] twice, once to get the overall number of tokens lemmatized without *się* (type *tworzyć*) and then to retrieve the occurrences of the equivalent tokens accompanied by *się* (type *tworzyć się*).[10] The outcome of the second query was subtracted from the first one, in order to roughly assess the frequency of the unit without *się*. However, obviously the second query did not match all the occurrences of verbs accompanied by lexical *się* (*tworzyć się*), since at the same time it retrieves an unknown number of occurrences with grammatical *się* relating to the lexical unit without *się* (*tworzyć*). Technically speaking: the second query yielded a number of false positives. This lack of precision cannot be removed without an inspection of every single occurrence, which is of course unfeasible. Thus, we performed a manual inspection of 28 ipfv. stems for which there exist both a lexeme with and without *się*. These stems were randomly selected from the database, on the condition that the *się*-verb had a frequency of at least 50 (in the PNC). For each of the selected verbs, again, a random selection of 50 concordance lines (= tokens) was drawn from the PNC (using the function SAMPLE in R). The concordances were classified manually to sort out false from true positives. Among these 28 verbs, 25 had a precision of 82% and higher, i.e 'lexical *się*' clearly predominated; only for two verbs 'grammatical *się*' turned out more dominant.

See the following table. Notably, the majority of these stems pairwise belong to the same aspect triplet in the database; this is indexed by the characters in square brackets. This coincidence, however, does not bear on the validity of the performed procedure; it only demonstrates that for certain triplets both ipfv. members (IPFV1 and IPFV2) reveal a comparably high frequency (which thus contributes to the token frequency of the entire triplet). On this issue see §4.3 in the main article.

[8] That is, with and without 'lexical *się'* as characterized above. An additional check was performed on the dictionary Doroszewski (1958-1968) to make sure that lexemes that only occurred either with or without *się* in the database were also accounted for, although their respective counterparts are registered by standard lexicography as separate lexemes (and could thus occur in corpora).

[9] The queries were made on the balanced version of the PNC, *KorBa* (for the 18th century) and two smaller corpora comprising the 19th century (*Pol_preWar* and *DiAsPol*; see References).

[10] The second query was formulated this way (STEM stands as a placeholder for the particular verb stems):

[orth=się] [base=STEM] | [base= STEM] [orth=by]? [pos=aglt]? [orth='mu|mi|ci|im|nam|jej']? [orth=się] .

In postposition *się* can only be separated from its stem by other enclitics (classified in PNC as "aglt" = "agglutinants"). As concerns *się* preceding its stem, some preparatory trials showed that *się* almost never was separated from its stem by other material, i.e. it almost exclusively occurred immediately before the stem. When the query included the possibility of more distance between the stem and preceding *się*, the recall did not improve much, but precision became much worse. This is why the query was restricted to *się* in immediate precedence to the stem and to the position following it.

verb: infinitive of forms marked with *się* in PNC	**precision rate**	
zadłużać się ‘get into debt’, IPFV2	100%	**lexical *się* clearly predominates**
zmieniać się ‘change’, IPFV2		
chylić się ‘bow’ (intr.), IPFV1 [c]		
dłużyć się ‘drag on’, IPFV1		
giąć się ‘bend’ (intr.), IPFV1 [g]		
powodzić się ‘prosper’, IPFV2 [a]		
przytulać się ‘cuddle up (to sb)’, IPFV2 [f]		
stapiać się ‘melt’ (intr.), IPFV2 [b]		
tulić się ‘cuddle’, IPFV1 [f]		
wychylać się ‘lean out’ (intr.), IPFV2 [c]		
wyginać się ‘bend’ (intr.), IPFV2 [g]		
zasklepiać się ‘isolate oneself’, IPFV2		
zwać się ‘be called’, IPFV1 [d]		
pokrywać się ‘overlap’, IPFV2	98%	
ciągnąć się ‘extend’ (intr.), IPFV1 [h]		
topić się ‘melt’ (intr.), IPFV1 [b]		
wieść się ‘be doing well’, IPFV1 [a]		
doskonalić się ‘improve oneself’, IPFV1	96%	
uwalniać się ‘free oneself, break free’, IPFV2		
wychowywać się ‘be brought up’, IPFV2	94%	
brać się ‘get involved, deal with’, IPFV1 [e]	92%	
nazywać się ‘be called’, IPFV2 [d]		
zaznaczać się ‘be(come) salient’, IPFV2	88%	
zabezpieczać się ‘protect oneself’, IPFV2	84%	
zabierać się ‘get going’, IPFV2 [e]	82%	
wyciągać się ‘stretch out’, IPFV2 [h]	46%	
obcinać się ‘get one’s hair cut’, IPFV2	20%	**grammatical *się* clearly predominates**
umieszczać się ‘be placed’, IPFV2	8%	

Precision rates of ‘lexical *się’* in Polish (sample of 28 verbs)

Beforehand, the occurrence of false negatives (i.e. the lack of accuracy of recall) had been estimated on the basis of 17 *reflexiva tantum*-verbs (like *smiać się* ‘laugh’, *bać się* ‘be afraid’). From these, 12 yielded a recall higher than 80%. This rate roughly equals the precision rates of the aforementioned random sample from the database. Thus, although we could test only small amounts of verbs – and for each pair of cognates with and without *się* there are a considerable number of factors that influence the accuracy of recall – the number of false negatives and false positives is similar and seems to cancel each other out. This allows us to consider the frequencies obtained from the aforementioned two queries (for verb forms with and without *się*) as reflecting more or less the “true” relation between verb lexemes with and without *się*. Still, we have to bear in mind that these figures are only a rough approximation. Put simply, while the sums of the frequencies of pairs of the type *tworzyć – tworzyć się* can be treated with high confidence, one has to remain cautious as for the shares of the respective *tworzyć*-like and *tworzyć się*-like stem in these sums.

Consequences for the database

As stated above, only for Polish could we provide a rough assessment of the repercussions which the reflexive marker has for the frequency counts of the units in the database.

Consequently, the frequency counts for entries with lexical *se/się/sja* have to be taken with caution, and to different extents.

As for Czech, we were unable to figure out a method to estimate the frequencies of triplets with *se*. We treated them in two ways. First, for triplets which, in the database, appear with and without the clitic *se*, e.g. *budit – probudit – probouzet* 'awaken, raise' and *budit se – probudit se – probouzet se* 'wake up', it was impossible to separately establish the frequencies of *budit* and *budit se*, and of *probouzet* and *probouzet se*, respectively. In such cases, the frequency of the stem without clitic *se* includes the frequencies of both stems, i.e. the frequencies are jointly assigned only to the stem without *se*. In studies for which frequency played a crucial role (see §4), such triplets were excluded. Second, for triplets which, in the database, appear only with *se*, e.g. *pálit se – spálit se – spalovat se* 'burn (down) [intr.]' (their transitive counterparts, without *se*, do not create triplets), the frequencies of the *se*-verb (*pálit se*, *spalovat se*), conversely, include the frequencies of the transitive equivalents (*pálit*, *spalovat*).

As concerns Polish, the results of the frequency counts for triplets that, in the database, occur both with and without 'lexical *się'*, are given for each of them according to the procedure described above. That is, no further calculations were made since the estimations for false positives and false negatives seem to cancel each other out. In triplets which entered the database only without or with 'lexical *się'*, regardless of whether the respective counterpart exists beyond the database, all occurrences in the corpus were attributed to this single verb. The (in)accuracy rate of these automatic counts should be comparable to the rate of pairs like *tworzyć—tworzyć się*, provided 'grammatical *się*', on average, can be considered as infrequent as with the 23 out of 26 randomly selected pairs of this type (see above), or, in other words, provided we may assume that only about 10% of verbs show a share of more than 80% of "grammatical *się*" in their token frequency. Again, this is but a rough estimation.

As for Russian, frequency counts for entries with the postfix *-sja* and without it are less prone to uncontrollable distortions, since the range of *-sja* in voice-like functions ('grammatical *-sja*') is much narrower than in the two West Slavic languages. Already this, in comparison to Polish and Czech, diminishes the potential for occurrences of *sja*-forms that do not operate on the verb's argument structure, the only serious remainder being *-sja* as marking the passive with ipfv. verbs. Admittedly, the proportion of discourse tokens in which ipfv. *sja*-forms are to be treated as the passive of transitive verbs is difficult to assess in general. However, Goto/Say (2009), who conducted the only systematic corpus-based study to date we know of, found that just about 3.3% of all tokens of *sja*-forms in the RNC can be counted as markers of the passive (of the respective transitive verbs).[11] We therefore left the figures of IPFV1 and IPFV2 stems in the database as we gained them from queries in the RNC[12], being aware of the general caveat.

References

Doroszewski, Witold. 1958-1968. *Wielki słownik języka polskiego*, t. 1-10. Warszawa: PWN.

Goto, Ksenija V. & Sergey S. Say. 2009. Častotnye xarakteristiki klassov russkix refleksivnyx glagolov. In: Kiseleva, Ksenija L., Vladimir A. Plungjan, Ekaterina V. Raxilina & Sergej G. Tatevosov (eds.). *Korpusnye issledovanija po russkoj grammatike*. Moskva: Probel-2000, 184–223.

Górski, Rafał L. 2008. *Diateza nacechowana w polszczyźnie (Studium korpusowe)*. Kraków: Lexis.

Holvoet, Axel. 2020. *The Middle Voice in Baltic*. Amsterdam, Philadelphia: Benjamins.

Kemmer, Suzanne. 1993. *The Middle Voice*. Amsterdam. Philadelphia: Benjamins.

Nedjalkov, Vladimir P. 2007. Overview of the research. Definitions of terms, framework, and related issues. In: Nedjalkov, Vladimir P. (ed.). *Typology of reciprocal constructions*, vol. I (ch. 1). Amsterdam, Philadelphia: Benjamins, 3-114.

[11] Their counts are based on the analysis of all approx. 10,000 *sja*-form tokens (participles excluded) gained from the belletristic subcorpus from the 1960s to the 2000s.

[12] We thank Oleg Bulatovs'kyj (L'viv) for his support in performing more complicated corpus searches.

Panevová, Jarmila, Markéta Lopatková & Václava Kettnerová. 2020. Reciproka ve slovníku a v syntaxi. In: Bílková, Jana, Ivana Kolářová & Miloslav Vondráček (eds.). *Lingvistika. Korpus. Empirie*. Praha: Ústav pro jazyk český, 63-70.

Rytel-Kuc, Danuta. 1990. *Niemieckie passivum i* man-*Sätze a ich przekład w języku czeskim i polskim*. Wrocław etc.: Ossolineum.

Wiemer, Björn. 1999. The light and the heavy form of the Polish reflexive pronoun and their role in diathesis. In: Böttger, Katharina, Markus Giger & Björn Wiemer (eds.). *Beiträge der Europäischen Slavistischen Linguistik (POLYSLAV)* 2. München: Sagner, 300-313.

Wiemer, Björn. 2007. Reciprocal and reflexive constructions in Polish. In: Nedjalkov, Vladimir P. (ed.). *Typology of reciprocal constructions*, vol. II (ch. 11). Amsterdam, Philadelphia: Benjamins, 514-559.

Wiemer, Björn (forthc.). Grammar of Voice. In: Bermel, Neil & Jan Fellerer (eds.). *Oxford Guides to the World's Languages. The Slavonic Languages*. Oxford University Press.