Language Resources and Evaluation (2021) 55:209–239
https://doi.org/10.1007/s10579-020-09511-7

PROJECT NOTES

Składnica: a constituency treebank of Polish harmonised with the *Walenty* valency dictionary

Marcin Woliński[1] · **Elżbieta Hajnicz**[1]

Accepted: 1 October 2020 / Published online: 21 February 2021

Abstract
This paper reports on the developments in three interrelated linguistic resources for Polish. The first is Świgra 2—a rule based constituency parser for Polish. The second is Składnica—a treebank built using Świgra 2. The third resource is valency dictionary Walenty, which became available when the work on the first two was already advanced. However, since the dictionary is much more comprehensive than the ad-hoc dictionary used previously with Świgra, a decision was made to switch the parser and the treebank to the new dictionary. The switch required several modifications to the Świgra 2 parser, including implementation of unlike coordination, introducing semantically motivated phrases, and non-standard case values. A semi-automated procedure to upgrade previously disambiguated trees in Składnica was required as well. Modifications introduced in the treebank during the upgrade included systematic changes of notation and resolving newly introduced ambiguities resulting from the use of the more detailed distinctions made in the dictionary. The procedure for confronting Składnica with the trees generated with the new version of the Świgra 2 parser using the Walenty dictionary allowed us to check all of these resources for consistency. This resulted in several corrections being introduced in both the treebank and the valency dictionary.

Keywords Treebank of Polish · Constituency parsing · Valency dictionary

1 Introduction

Treebanks—corpora annotated with syntactic information—have an established position as an important tool both for linguistic inquiries and for machine learning.

✉ Marcin Woliński
wolinski@ipipan.waw.pl

Elżbieta Hajnicz
hajnicz@ipipan.waw.pl

[1] Institute of Computer Science, Polish Academy of Sciences, Jana Kazimierza 5, 01-248 Warsaw, Poland

However, treebanks of Polish are not yet abundant. Składnica (Woliński et al. 2011) is the first treebank of Polish of a considerable size. It is a constituency treebank consisting of trees generated by the Świgra 2 parser of Polish and then manually disambiguated and validated.

Besides Składnica, publicly available treebanks of Polish include: a dependency treebank of Polish (Wróblewska 2014), which includes converted Składnica plus trees prepared manually, and an LFG structure-bank prepared using the grammar POLFIE (Patejuk and Przepiórkowski 2014). Each of these resources is also available in Universal Dependencies form (Wróblewska 2018; Przepiórkowski and Patejuk 2020).

Walenty is currently the largest valency dictionary of Polish (Przepiórkowski et al. 2014b). Moreover, it is available in a machine-readable format and it is the most advanced in linguistic features, in particular it has a phraseological component. The availability of a large independently maintained valency dictionary is a game changer for Świgra 2 and Składnica. Therefore, deploying Walenty was an obvious choice for the further development of Składnica.

Both Składnica and Walenty are based on the National Corpus of Polish (in Polish: *Narodowy Korpus Języka Polskiego*, NKJP) (Lewandowska-Tomaszczyk et al. 2013; Przepiórkowski et al. 2012). Składnica is built from utterances extracted from the NKJP, whereas every syntactic schema of Walenty is illustrated with example sentences drawn from the NKJP.

Valency dictionaries, especially semantic ones, are often associated with a corpus of examples illustrating particular valency frames. This is the case with FrameNet (Fillmore et al. 2003) and VerbNet (Kipper et al. 2008). Unfortunately, these corpora are not treebanks, even though phrases are marked according to their semantic roles in FrameNet's exemplary sentences.

An example of a treebank coupled with a valency dictionary is PropBank (Kingsbury and Palmer 2002; Palmer et al. 2005). Its core part is composed of the Wall Street Journal portion of the Penn Treebank (Marcus et al. 1993) augmented with predicate-argument structures comprising the valency dictionary part of the project. In contrast to other dictionaries, verbs' semantic arguments have no labels, but are simply numbered, from 0 up to 6.

Starting with its 2.0 edition, the Prague Dependency Treebank (Böhmová et al. 2003; Hajič 2005, PDT) contains tectogrammatical annotation including deep syntax synchronised with the PDT-Vallex valency lexicon (Urešová 2009). PDT-Vallex covers only the predicates occurring in the PDT and contains only valency frames attested in the treebank. The representation of valency is very detailed. In particular, the strength of phraseological formalisms used in PDT-Vallex and in Walenty are similar (cf. Przepiórkowski et al. 2017).

It is worth noting that another Czech valency dictionary, Vallex (Žabokrtský and Lopatková 2007; Kettnerová et al. 2012) shares with the PDT-Vallex common theoretical underpinnings anchored in the Functional Generative Description (FGD) theory (Sgall et al. 1986). However, it aims at providing complete descriptions of a possibly large number of lexemes with less detailed information.

In the paper, we describe the procedure for adapting Składnica to the new valency dictionary and its results. The procedure was to a large extent automatic. However, the differences between the resources made it necessary to correct some parse trees

manually and to resolve new ambiguities introduced due to a more detailed taxonomy of arguments in Walenty compared to the old dictionary. We present the method of automatic mapping and the problematic cases that needed manual intervention.

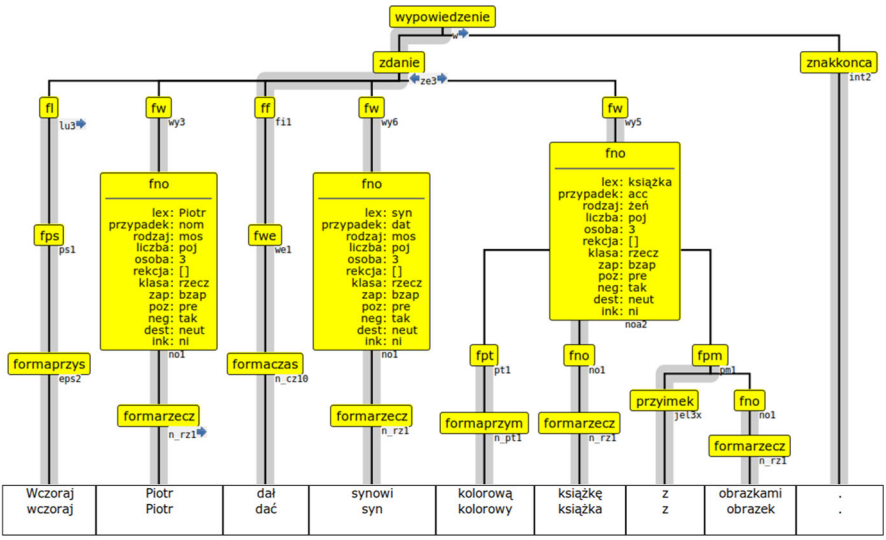
The article is organised as follows. First, we describe the resources—the parser Świgra 2 (Sect. 2) and the syntactic structures it generates (Sect. 3), and the treebank Składnica (Sect. 4), and finally the valency dictionary Walenty (Sect. 5). Next, we analyse changes in the Świgra 2 parser needed to deploy Walenty (Sect. 6) and show some constructions which can be analysed thanks to the change of dictionary (Sects. 7 and 8). Finally, we discuss the process of upgrading the Składnica treebank (Sect. 9) and evaluate the resulting resource (Sect. 10).

2 The parser Świgra 2

Świgra is a DCG (Pereira and Warren 1980) rule-based constituency parser of Polish. The grammar used by Świgra stems from Świdziński's grammar (Świdziński 1992), whose implementation was called Świgra 1 (Woliński 2004). For the new version, called Świgra 2, the grammar has been considerably restructured (Woliński 2019; Świdziński and Woliński 2010). The trees generated by Świgra 2 are much simpler and more intuitive but still capture all essential information present in the structures generated by the old grammar. In particular, binary branching of trees was abandoned for all types of syntactic constructions, whereas a natural n-ary structure has been proposed. As a result, the trees frequently have nodes of high arity but their height is much lower than in Świdziński's grammar. The number of non-terminal categories has been greatly reduced. For example, there is just one category **zdanie** for sentences/clauses and one category **fno** for nominal phrases, while Świdziński's grammar used 5 different units for sentences and 6 units for various sub-types of nominal phrases.

The grammar of Świgra 2 was also extended to cover many constructions not considered in Świdziński's grammar (Woliński 2019). First of all, coordinated structures are now allowed in all types of constructions: sentences and various phrases. The description of nominal phrases became more advanced: numerals were introduced as possible constituents and apposition is now a possible means of joining nominal phrases. We have also described the possibility of particles to modify phrases of various categories (for that, a taxonomy of particles had to be implemented). A special type of coordinated nominal structures has been described where the phrase as a whole does not share the characteristics of any of its constituents (e.g. two coordinated phrases in singular act as a plural phrase). Another new feature of the grammar is sentence-like constructions that lack a verbal form as their centre.

From the very beginning, Świgra has been using a valency dictionary for verbs, which gets consulted from grammatical rules describing verbal phrases and clauses. In Polish, most argument types are optional, so often only a subset of a syntactic schema is realised. In particular, Świgra is careful to generate only one tree when subsets of several schemata can be used to analyse a given sentence. The mechanics of filling the valency slots are described in more detail in the paper (Woliński 2015), including the version used for parsing with Walenty.



Wczoraj Piotr dał synowi kolorową książkę z obrazkami.
yesterday Peter gave son colourful book with pictures

'Yesterday Peter gave [his] son a colourful illustrated book.'

Fig. 1 An example of a constituency tree generated by Świgra

Two advanced features of the Świgra 2 grammar will be discussed in separate sections. The first one is coordination of arguments of different types (Sect. 7), which is a new feature of Walenty that required changes in the implementation of valency in Świgra. This change also made sharing of arguments between predicates possible. It was also a good occasion to implement another advanced feature, namely, parsing of some common discontinuous Polish constructions (Sect. 8).

The parser is available for download at the address http://zil.ipipan.waw.pl/%C5%9Awigra, while an on-line version can be accessed at http://swigra.nlp.ipipan.waw.pl/.

3 Syntactic trees of Świgra and Składnica

Świgra uses constituency trees as a representation of syntactic structures. An example of such a tree is shown in Fig. 1. Leaves of the tree correspond to terminals (forms and lemmas shown in the boxes at the bottom of the picture). Internal nodes of the tree correspond to non-terminals of the grammar. They are represented by the name of the non-terminal category in the figure. The labels use abbreviations of Polish names, which are explained in Table 1. As is natural for a unification grammar, nodes of the tree carry sets of attribute-value pairs specifying their syntactic features (e.g. features of nominal arguments **fno** are shown explicitly in Fig. 1). The children of a given node are its constituents, as determined by the used rule of the grammar.

The non-terminals of the grammar conceptually fall into several types or layers in the trees (Świdziński and Woliński 2010). From the bottom up, these are:

Table 1 Non-terminal categories used in Świgra and Składnica (selection)

Syntactic forms	
formaczas	Verbal form
formarzecz	Nominal form
formaprzym	Adjectival form
formaprzys	Adverbial form
zaimrzecz	Nominal pronoun
zaimos	Personal pronoun
spójnik	Conjunction
przyimek	Preposition
znakkonca	Final punctuation
Constituent phrases	
fwe	Verbal phrase
fno	Nominal phrase
fpt	Adjectival phrase
fps	Adverbial phrase
fpm	Prepositional-nominal phrase
fzd	Clausal phrase
ff	Finite phrase (an **fwe** that can constitute a clause)
Valency phrases	
fw	Argument
fl	Adjunct
Clauses	
zdanie	Clause/sentence
wypowiedzenie	Utterance

1. *Syntactic forms*, which are the syntactic counterpart of inflectional forms (terminals of the grammar). Typical examples are the units **formaczas**, **formarzecz**, **formaprzym**, and **przyimek** in Fig. 1. However, units of this level can also represent multi-token verbal forms (e.g., analytical future forms of verbs like *będziemy mogli* '[we] will be able') and other cases where one form, from the syntactic viewpoint, corresponds to several tokens in the NKJP tagset, e.g. two-word prepositions *wraz z* 'together with' and adverbs *po ciemku* 'in the dark'.
2. *Constituent phrases* are used to describe the attachment of various dependants to verbal, nominal, adjectival, and adverbial heads. Also at this level, prepositional-nominal phrases and subordinate clauses are formed. Constituent phrases can also be coordinated structures (with a conjunction as a head).
3. *Valency phrases*, as proposed by Świdziński (1992), denote functions played by constituent phrases. These differentiate dependants into argument phrases **fw** ('required phrases' according to the terminology adopted by Świdziński) and

adjunct phrases **fl** ('free phrases'[1]). Thanks to this layer, the basic shape of the valency structure becomes visible in the tree.

4. The fourth layer comprises *clauses* represented by the non-terminal **zdanie**. Simple clauses consist of a finite phrase **ff** and valency phrases. Coordinate clauses, based upon a conjunction as their head, have other clauses as their constituents.

For example, in Fig. 1 the word *Piotr* is interpreted at the first level as a syntactic noun **formarzecz**, which is treated as a constituent nominal phrase **fno**, playing a valency role of an argument **fw**, which becomes a constituent of a clause **zdanie**. This branch of the tree goes through the four levels in sequential order. The layers can get tangled, e.g., when a relative clause (level 4) becomes a constituent of a nominal phrase (level 2).

An important feature of Składnica trees is the fact that one of the constituents is labelled as the syntactic head (marked in the picture with a thick grey background around an edge), which allows constituency trees to be converted into dependency trees. Such conversion has in fact been performed resulting in a dependency version of Składnica (Wróblewska and Woliński 2012), later on also converted to Universal Dependencies (Seddah et al. 2013).[2]

The link between phrases and syntactic schemata is provided by an attribute of argument phrases **fw** named tfw—the 'type of governed phrase'. This attribute shows the type of phrase according to the notation used in the valency dictionary. For example, the three arguments in Fig. 1 have, respectively, tfw=subj(np(nom)) for *Peter*, which is a nominal phrase in the nominative, tfw=np(dat) for *synowi*, which is in the dative, and tfw=np(accgen) for *kolorową książkę z obrazkami*, which is in the structural case (cf. Sect. 5). Since the first phrase is marked as the subject subj, the rules of the grammar ensure the agreement of the person, number, and gender of this phrase and the finite head **ff**. In the example, the verb *dał* can be interpreted as singular of any masculine sub-gender. But agreement with the subject *Piotr* limits the gender to masculine personal mos.[3]

The part of the valency schema realised for the given predicate becomes the value of the attribute rekcja 'valency' of the respective phrase. In the example, the value of rekcja assigned to **ff** and **fwe** corresponding to *dał* is equal [subj(np(nom)), np(dat), np(accgen)]. As mentioned before, this value can be a common subset of several schemata for the given verb (in the degenerate case, even of all schemata for the verb). For this reason, no link is provided to a particular schema in the dictionary.

[1] In fact, Świdziński's free phrases are a broader term than adjunct. They include also elements which are present in a sentence but which are not part of the structure, e.g. vocative phrases which are not dependants of a verb.

[2] The conversion was performed on an earlier version of Składnica. For the newest version the conversion procedure will have to be updated.

[3] This is achieved with a somewhat counter-intuitive technical trick: most attributes of **fw** and **fl** mimic the attributes of the finite head of the clause. This allows to confront selected attributes of the given dependant with the attributes of the head. This mechanism is used also for other agreements, e.g, some nominal adjuncts **fl** in the vocative are required to agree with the verb.

4 The treebank Składnica

Składnica is a treebank of Polish that was conceived as a means to aid the development of the grammar for Świgra 2 and to test its corpus coverage (Woliński et al. 2011).

The texts included in Składnica were sampled from the one million word sub-corpus NKJP1M of the National Corpus of Polish. NKJP1M is very convenient for such work, since it has been manually annotated on the morphological level. Every token of the corpus has an unambiguous validated morphological interpretation. For Składnica, we have extracted samples from NKJP1M, which are a few sentences long each and sum up to 20,000 sentences.

The text is parsed with Świgra, which results in ambiguous parse forests. The forests are manually disambiguated and validated using a web based system named Dendrarium (Woliński 2010). During the process, annotators choose interpretations for ambiguous nodes of the forest. Each sentence is presented to two annotators independently. If there are conflicts in annotation, an adjudicator steps in. It was assumed as a construction rule for the treebank that all accepted trees have to be actually generated by the parser. The treebank annotators are not allowed to modify trees in any way nor to provide trees for sentences rejected by the parser. If a tree is finally selected, the annotator has to check whether it is consistent with the annotation guidelines. This is not always the case: even if the parser has succeeded in fitting the sentence to some structure it knows, the sentence may in fact be an example of a language construct not covered by the grammar. In such a case, the tree is rejected and a comment describing the reason for rejection is obligatory. The grammar is corrected and offending sentences parsed anew. This leads to an iterative development of the grammar and the treebank. The grammar feeds the treebank and the treebank documents the coverage of the grammar. Thus, an important feature of Dendrarium is a module that transfers trees accepted by the annotators to new forests generated by changed versions of the grammar.

The first version of the treebank, named Składnica 0.5, was developed in the years 2009–2011 in a Polish Ministry of Science financed project N N104 224735. In this project, trees for 8227 sentences have been accepted by annotators (41.1% of 20,000 sentences). The inter-annotator agreement was 88%, measured for whole sentences. The rejected sentences were classified, and the most common reason for rejection turned out to be the presence of related speech (*oratio recta*).

Składnica was further developed in the following years. During the process, the rules of the grammar were gradually improved based on the obtained classification of problematic sentences. Finally, the valency dictionary used by the parser was replaced with Walenty, leading to the version described in this article.

Składnica is available for download in the form of XML files at the address http://zil.ipipan.waw.pl/Sk%C5%82adnica.

5 Valency dictionary Walenty

A valency dictionary specifies what types of arguments are possible for a given predicate. The need for such information is most obvious for verbs, which differ widely in possible arguments, e.g., some Polish verbs allow for a complement in the form of

a verbal phrase in the infinitive and others do not. Other classes of predicates have mostly typical dependants—for instance, adjectival and prepositional-nominal phrases for nouns. Infinitival dependants do not occur with nouns and are very rare for adjectives. Yet providing valency information for other classes of predicates is also useful, especially when differentiating arguments and adjuncts is involved.

Initially, the Świgra parser used a valency dictionary based on (Świdziński 1994). This dictionary was extended when the Składnica treebank was built. Its version released with Składnica 0.5 consisted of 6400 schemata for 1450 Polish verbs, covering about 75% of verb occurrences in the 1 million tokens manually annotated subcorpus of the NKJP (Woliński et al. 2011). The dictionary only contained verbs.

Later, this dictionary became a seed for a new one, which is currently being developed at the Institute of Computer Science of the Polish Academy of Sciences (ICS PAS). The new dictionary, called Walenty, is a comprehensive valency dictionary of Polish based on corpus data (Hajnicz et al. 2016a, b; Przepiórkowski et al. 2014a, b, c). After several years of development, Walenty contains 101,500 schemata for 18,250 predicates, which include about 13,000 verbs, 4000 nouns and 1100 adjectives and adverbs. Walenty covers 99.8% of occurrences of verbal forms in the 300 million word balanced sub-corpus of the NKJP. Moreover, Walenty is much richer in linguistic information than the original dictionary of the Świgra parser. Among other features, it describes syntactic control and raising and contains a rich phraseological component.

Walenty consists of two layers. On the syntactic level of Walenty, valency is expressed in terms of syntactic types of phrases (e.g., nominal phrase, verbal phrase) and their grammatical features (e.g., case, aspect). Phrases of specified types fill syntactic positions, which comprise syntactic schemata. The second level describes semantics by coupling syntactic schemata with semantic frames consisting of arguments specified as semantic roles and their selectional preferences (Hajnicz et al. 2016a).

In this paper we are only concerned with the syntactic layer. Thus we will consider a dictionary entry for a predicate to be a set of *valency schemata*[4]. Each schema is a set of *syntactic positions*, which can be realised by arguments of specified *phrase types*.

5.1 Phrase types

Phrase type specification in Walenty describes the kind of allowed syntactic construction and required grammatical features. The list of phrase types includes nominal phrases (np), adjectival phrases (adjp), prepositional phrases (nominal prepnp and adjectival prepadjp), infinitival phrases (infp), clausal phrases[5] (cp), clausal phrases with a nominal correlate (ncp) and clausal phrases with a prepositional-nominal correlate (prepncp). The phrase types have several attributes specifying their grammatical features. Table 2 presents attributes governed by the predicate for each phrase type and lists possible values of particular attributes. For the complete specification of available phrase types, see (Hajnicz et al. 2016b).

[4] We use the term *schema* at the syntactic level and the term *frame* at the semantic level.

[5] This non-traditional term denotes a subordinate clause together with a complementizer that introduces the clause.

Table 2 Selected phrase types and their attributes in Walenty

Type	Attributes
np(CASE)	
adjp(CASE)	
prepnp(PREPOSITION, CASE)	
prepadjp(PREPOSITION, CASE)	
infp(ASPECT)	
cp(CTYPE)	
ncp(CTYPE, CASE)	
prepncp(CTYPE, PREPOSITION, CASE)	
xp(SEMTYPE)	

Attribute	Possible values
CASE	nom, gen, dat, acc, inst, loc, accgen, part, agr, pred
PREPOSITION	any preposition
ASPECT	imperf, perf
CTYPE	aż, gdy, jak, jakoby, jeśli, kiedy, że, żeby, int, rel
SEMTYPE	locat, abl, adl, perl, temp, dur, mod, cause, dest, instr

Grammatical case is governed by the predicate for nominal, adjectival and prepositional phrases. Similarly, the predicate governs the case of nominal (ncp) and prepositional-nominal (prepncp) correlates. Apart from six usual case values, some special ones are used in the dictionary. The most important is the so-called structural case, i.e. the case whose morphological realisation depends on the syntactic context. Structural case is used to specify nominal phrases underlying the genitive of negation. In Świgra and Składnica, we denote this structural case with the mnemonic symbol np(accgen), since this type of phrase is realised in the accusative or in the genitive, depending on whether the predicate is negated or not.[6] This can be illustrated with a simple schema for the verb *jeść* 'to eat' (imperfect):

(1)

subj	obj	
np(nom)	np(accgen)	prepnp(na,acc)

The schema comprises a subject position, a nominal object position in the structural case and a position for a prepnp phrase type containing the preposition *na* with an accusative complement. This schema can be applied to an affirmative sentence (2) and a negated sentence (3). Observe that the object *mięso* 'meat' in (2) is in the accusative, whereas *owoców* 'fruit' in (3) is in the genitive case.

(2) *Codziennie* *jem* *mięso* *na* *śniadanie,* *obiad* *i*
every day eat.SG.PRES meat.ACC.SG for breakfast.ACC.SG lunch.ACC.SG and
kolację.
dinner.ACC.SG

'[I] eat meat for breakfast, lunch and dinner every day.'

[6] In Walenty, structural case is used also for subjects according to Przepiórkowski's controversial concept of numeral subjects in Polish being in the accusative (Przepiórkowski 2004). In Świgra nominal phrases in the subject position are considered to be simply np(nom).

(3) *Dzieci* *przez* *długie* *miesiące* *nie* *jedzą* *owoców.*
children.NOM.PL for long.ACC.PL month.ACC.PL not eat.PL.PRES fruit.GEN.PL

'Children do not eat fruit for long months.'

The other non-standard values for case are used when the grammatical case depends on the predicate in some convoluted manner. The symbol part represents the so called partitive case, agr denotes agreement of a dependant with the head in phraseological schemata, and the so called predicative case pred is used for adjectives in the predicative position (Przepiórkowski et al. 2014a).

5.2 Clausal phrases

The kind of clausal phrase cp is typically determined by specifying the complementizer introducing the clause, e.g., *jeśli* 'if', *kiedy* 'when', *że* 'that' or *żeby* 'in order to'. Two types are not introduced by a complementizer. These phrases are bare clauses of a specific type: relative clauses cp(rel), which must contain a relative pronoun in the initial constituent, and interrogative clauses cp(int), whose initial constituent has to be interrogatory.

A simple schema (4) for the verb *podejrzewać* 'suspect' containing an interrogative clausal phrase is exemplified by sentence (5) with a subordinate interrogative clause *kim był denat* 'who the deceased was' introduced by the interrogative pronoun *kto* 'who'.

(4)

subj	
np(nom)	cp(int)

(5) *Policja* *podejrzewa,* *kim* *był* *denat.*
police.NOM.SG suspect.SG.PRES who.INST.SG be.SG.PAST deceased.NOM.SG

'Police suspect who the deceased was.'

Clausal phrases can appear with a nominal (ncp) or a prepositional (prepncp) correlate. The correlate is a form of the pronoun *to* 'this' in a governed case, optionally appearing after a preposition. These constitute separate phrase types, since they are not (always) interchangeable with cp. Sentence (6) contains a clause *tym, że pokazuje w każdej piosence nieco inną siebie* 'by showing herself from a slightly different side' of type ncp(inst,że), which is introduced by the nominal correlate in the instrumental followed by the complemetizer *że* 'that'. Sentence (7) contains a clause *o tym, by jeździć bezpiecznie* 'to drive safely' of type prepncp(o, loc, żeby) composed of the preposition *o* 'about' governing correlate *tym* 'this' in the locative followed by complementizer *żeby* 'in order to'. The schemata used in these examples will be discussed on page 28,

as they use coordination, cf. (10) for the verb *urzekać* 'to charm' and (13) for *pamiątać* 'to remember'.

(6) *Na obu płytach Banaszak urzeka tym, że*
on both.LOC disc.PL.LOC surname.SG.NOM charm.SG.PRES this.SG.INST that
pokazuje w każdej piosence nieco inną siebie.
captivate.SG.PRES in every.SG.LOC song.SG.LOC a little different.SG.AAC self.SG.ACC

'On both albums, Banaszak charms [us] by showing herself from a slightly different side.'

(7) *Niestety, nie wszyscy pamiętają o tym, by*
Unfortunately not everyone.PL.NOM remember.PL.PRES about this.SG.LOC to
jeździć bezpiecznie.
drive.INF safely

'Unfortunately, not everyone remembers to drive safely.'

5.3 Semantically motivated phrases

Walenty provides semantic classification of some adverbial-like arguments (e.g., ablative and adlative), denoted as xp(...). Such valency positions can be filled mainly with adverbs and prepositional phrases. The attribute of xp specifies a semantically motivated set of allowed realisations. For example xp(abl)—ablative phrase, marking the departure point of a motion—can be realised (among others) by adverbs *stąd* 'from here', *znikąd* 'out of nowhere', or prepnp(z,gen)—phrases with the preposition *z* 'from'. Adlative phrases xp(adl) denote point of arrival: *tutaj* 'here', *naprzód* 'forward', prepnp(do,gen)—*do* 'towards', complex preposition comprepnp(w kierunku) 'in the direction of', or even clauses, e.g. cp(rel[dokąd; gdzie])—a relative clause limited to two relative pronouns *dokąd* 'where to' and *gdzie* 'where'. The lists of allowed xp realisations are stored separately; their identifiers are used in schemata. In total, there are 10 specific subtypes of xp—expressing time, duration, place, starting or ending point, path, tool, manner, cause, or aim, cf. Table 2.

Ablative, adlative and perlative phrases are typical for verbs of movement. Below we present a schema (8) of the verb *maszerować* 'to march', illustrated by sentence (9), where the ablative phrase is realised by a prepositional phrase *z domu* 'from home', the adlative phrase—by a prepositional phrase *do szkoły* 'to school', and the perlative phrase—by a nominal phrase in the instrumental *niebezpieczną ulicą* 'dangerous street'.

(8)

subj			
np(nom)	xp(abl)	xp(adl)	xp(perl)

(9) *Dzieci* *maszerują* *niebezpieczną* *ulicą* *z* *domu*
children.NOM.PL march.PL.PRES dangerous.INST.SG street.INST.SG from home.GEN.SG
do *szkoły.*
to school.GEN.SG

'Children walk along the dangerous street from home to school.'

5.4 Syntactic positions

As can be seen in the previous examples, two positions are labelled in syntactic schemata: the subject subj (the nominal argument in this position influences morphological features of the finite verb) and the passivable object obj (the argument in this position turns into a subject in the passive voice; the presence of this position signals that passive voice is possible).

Walenty is explicit about what counts as a single syntactic position, and it employs the coordination test to resolve doubts in this respect: if two phrases can be coordinated in the same sentence then they are different realisations of the same position and they are listed in the same schema as alternative realisations for the given position. For instance, the sentence (11) contains two coordinated phrases np(inst): *solidność* 'solidity' and *pracowitość* 'diligence' and the clause with the nominal correlate ncp(inst,że) *tym, że na wszystko miała sposób* 'that she has a solution for everything'. The schema used to parse this sentence is (10).

(10)

subj	obj	
np(nom)	np(accgen)	np(inst) ncp(inst,int) ncp(inst,że)

(11) *Urzekała* *mnie* *solidnością,* *pracowitością* *i* *tym,* *że*
charm.SG.PAST I.ACC.SG solidity.INST.SG diligence.INST.SG and this.INST.SG that
na *wszystko* *miała* *sposób.*
for.ACC everything.ACC.SG have.SG.PAST solution.ACC.SG

'[She] charmed me with [her] solidity, diligence and the fact that she had a solution for everything.'

Sentence (12) is an example of two coordinated clauses with a prepositional correlate with the preposition *o*—prepncp(o, loc, że) (*o tym, że się ciężko pracuje* 'that sb works hard') and prepncp(o, loc, int) (*o tym, jaka jest sytuacja innych ludzi* 'about what the situation of other people is').

(12) *Otóż trzeba pamiętać nie tylko o tym, że się*
well must.PRES remember.INF not only about.LOC this.LOC.SG that refl
ciężko pracuje, ale także o tym, jaka jest
hard work.SG.PRES but also about.LOC this.LOC.SG what.NOM.SG be.SG.PRES
sytuacja innych ludzi.
situation.NOM.SG other.GEN.PL people.GEN.PL

'Well, you have to remember not only that you work hard, but also what the situation of other people is.'

The following schema can be used to analyse this sentence:

(13)

subj	
np(nom)	prepnp(o,loc) prepncp(o,loc,int) prepncp(o,loc,że) prepncp(o,loc,żeby)

Coordination is the main reason to allow clausal phrases cp in the subject position. Let us look at sentence (14) with a clausal argument *że zapomniałam, jak wyglądasz* 'that [I] forgot what [you] looked like'. The sentence could be modified and extended into (15) in which the clause is coordinated with a nominal subject *Piotr i że zapomniałam, jak wygląda* 'Peter and that [I] forgot what [he] looked like', which is an argument to assume that the cp(że) clause is a subject in (14). The respective schema for the verb *śnić* 'to dream' is shown as (16).

(14) *Śniło mi się, że zapomniałam, jak wyglą dasz.*
dream.SG.PAST I.DAT refl that forget.SG.PAST how look.SG.PRES

'I dreamed that [I] forgot what [you] looked like.'

(15) *Śnił mi się Piotr i że zapomniałam, jak wygląda.*
dream.SG.PAST I.DAT refl Peter.NOM.SG and that forget.SG.PAST how look.SG.PRES

'I dreamed about Peter and that [I] forgot what [he] looked like.'

(16)

subj	
np(nom) ncp(nom,że) cp(że)	np(dat)

Other cases of the so-called unlike coordination are discussed in Sect. 7.

5.5 Syntactic schemata

In Walenty, due to the free word order of Polish, the order of positions within a schema and the order of argument types within a position is not important.

Valency schemata given by Walenty are maximal—the dictionary does not list possible sub-schemata of a given schema. In Polish, most arguments are optional. In particular, subjects are often omitted. It is also possible to omit a direct object. A sentence with a missing direct object remains grammatical, but is usually semantically incomplete. Thus, transitivity is a much less sharp classification of Polish verbs than it is for English.

Only phraseological elements are strictly obligatory in Walenty. A schema with such elements cannot be applied if the phraseological arguments are missing in the sentence.

5.6 Phraseology

Walenty includes a rich phraseology component, implementing a detailed notation for various types of idiomatic arguments, from completely fixed (given as a string) to almost freely modifiable—in a recursive way (Przepiórkowski et al. 2014a, 2017; Hajnicz et al. 2016b). The dictionary aims at a precise representation of the structure of lexicalised arguments. For instance, schema (17) represents a phraseological construction *czuć się na siłach* 'to feel fit to do sth'. The idiomatic expression as a whole opens a position for infinitival phrase infp(_), whereas the verb *czuć się* 'to feel' itself does not. The type of a lexicalised phrase is denoted as lex with the first attribute specifying the syntactic type of the phrase (here prepnp(na,loc)). The type determines the other attributes. In this example, the phrase is required to contain a nominal phrase with the lexical head *siła* 'strength' in the plural pl, no modifiers are allowed in this phrase (natr). The construction is illustrated by sentence (18), where the infinitival phrase is *składać zeznania* 'to give testimony'. Note that this is an idiomatic expression as well, meaning 'testify' (in Polish: *zeznawać*).

(17)

subj		
np(nom)	infp(_)	lex(prepnp(na,loc),pl,'siła',natr)

(18) *Nie czuła się na siłach składać zeznania.*
not feel.SG.PAST refl on strength.LOC.PL give.INF testimony.ACC.PL

'She didn't feel strong enough to testify.'

This notation is also used to define so called compound prepositions comprepnp. These are typically prepositional-nominal phrases which from the valency point of view act as simple prepositions—they have an argument, typically a nominal phrase in genitive np(gen) (but a clause with a nominal correlate ncp(...,gen) is also frequent). For example, comprepnp(w kierunku) 'in [the] direction of' occurs directly in some

schemata and is a possible realisation of xp(adl) and xp(dest). For details of the internal structure notation, see (Hajnicz et al. 2016b; Przepiórkowski et al. 2017).

6 Adapting Świgra to Walenty

Adopting Walenty was a rather obvious decision in the development of Świgra but it meant that some changes needed to be introduced in the parser and in the grammar to adapt to a different format and to take advantage of the more detailed description. Simple changes included translating the symbols used in the old dictionary, which were based on Polish abbreviations for values of grammatical categories, to those used in Walenty (Latin/English based).

The most fundamental difference between the dictionaries is in the form of syntactic schemata. In the old dictionary, a schema is a flat list of phrase types which can be realised in a sentence. In Walenty, a schema is a list of positions, each of which is a set of alternative phrase types. To adapt to this difference the internal representation of schemata in the parser had to be redesigned. With long schemata and multiple partial matches, the use of Walenty can be quite complicated, which means an efficient way of using schemata in the parser had to be developed (Woliński 2015). This also made it possible to implement coordination within syntactic positions (Sect. 7) and, mostly as a by-product, to describe some discontinuous structures (Sect. 8).

To use Walenty's non-verbal schemata, the mechanism for filling syntactic positions was also introduced in rules defining nominal phrases (including those based on gerunds) and adjectival phrases (including adjectival participles).

To use lexicalised schemata such as (17), it was necessary to make the lemma of the lexical head of each phrase available. In DCG, information is only available locally—a grammar rule can only access the category and the information available as attributes of a given node. So, it was necessary to add attributes that carry the information on the lexical head along the 'head branch' of each subtree. With these changes, Świgra now uses phraseological schemata of Walenty (although the complete analysis of embedded modifiers of lexicalised items is not performed).

Semantically motivated phrase types xp also had to be implemented. The old dictionary uses a much less precise general type advp, so respective rules had to be replaced with ones defining the possible xp subtypes. This was easy, since all the necessary realisations were already covered by the grammar, they only get classified differently.

Walenty uses a broader concept of the subject than was used in the old dictionary. The label subj is applied not only to nominal phrases in the nominative, but also to some other phrases which can get coordinated with a nominal phrase, e.g. cp(że) in example (14). We decided to interpret subjects in the same way in Świgra, which means new rules had to be added for those realisations. As a result, much fewer verbs are inherently subjectless in this new interpretation.

New grammar rules had also been added to implement special types of arguments present in Walenty, e.g. complex prepositions.

7 Unlike coordination and argument sharing

As explained earlier, a position in a syntactic schema in Walenty is a set of phrase type specifications. The types specify alternative realisations of the given position. However, the fact that they are listed within a single position also means that arguments of these types can get coordinated. This is so called unlike coordination, as opposed to simple coordination where an argument is realised by a coordinated phrase of a single type. For example, schema (19) specifies the nominal phrase in the structural case np(accgen) as one of the possible realisations of the object position for the verb OKREŚLIĆ 'to determine'.

(19)

subj	obj		
np(nom)	np(accgen) cp(int) cp(że) ncp(accgen,int) ncp(accgen,że)	prepnp(w,loc)	xp(mod)

This licenses the following Polish sentence with simple coordination:

(20) *Jan określił rodzaj infekcji i właściwą kurację.*
John determined type infection and appropriate treatment

'John has determined the type of infection and the appropriate treatment.'

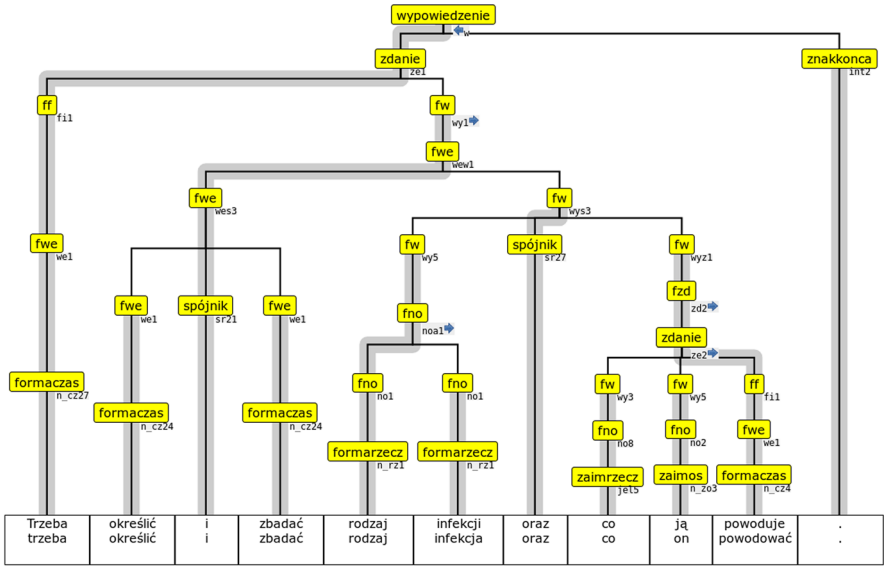
However, the same position contains specification cp(int), so it is possible to coordinate an interrogative clause with an np(accgen):

(21) *Jan określił rodzaj infekcji i co ją powoduje.*
John determined type infection and what it causes

'John has determined the type of infection and what causes it.'

In this sentence, a nominal phrase *rodzaj infekcji* 'type of infection' gets coordinated with a clause *co ją powoduje* 'what causes it'. If the coordination was not possible, separate schemata with respective phrase types would be given.

For such phrases, a problem emerges: what category to assign to a coordinated phrase consisting of a nominal phrase and a clause. Should it be called nominal, a clause, or some special type? In Świgra, such coordinated phrases are formed only to become arguments, so we took a rather elegant solution: such type of coordination happens on the level of argument phrases **fw**. So, the coordination in example (20) is covered by a new rule in the grammar stating that an argument phrase **fw** of type np(accgen) can get coordinated with **fw** of type cp(int) forming a new **fw** of a complex type.



Trzeba określić i zbadać rodzaj infekcji oraz co ją powoduje.
must determine and study kind infection and what it causes

'It's necessary to determine and confirm the kind of infection and what causes it.'

Fig. 2 A Świgra parse tree with unlike coordination and argument sharing

A complete structure for a sentence with this kind of coordination can be seen in Fig. 2. The phrase *rodzaj infekcji* 'kind of infection' is analysed as a nominal phrase **fno** in the accusative, which turns into an argument **fw** of type np(accgen). The clause *co ją powoduje* 'what causes it' becomes a clausal phrase **fzd** of type int—interrogative (*co* 'what' is an interrogative pronoun), and then an argument **fw** of type cp(int). These arguments get coordinated to become a phrase **fw** of type [np(accgen),cp(int)]. A mechanism was introduced that checks that this composite type is a subset of the appropriate position in some schema for the given verb. As can be seen, it is the case with the shown schemata for the verbs OKREŚLIĆ 'determine' and ZBADAĆ 'study' (the latter schema given as (22)).

(22)

subj	obj	
np(nom)	np(accgen) ncp(accgen,int) cp(int)	xp(instr)

What makes the example even more interesting, the two verbs are also coordinated and form a complex verbal phrase *określić i zbadać*. Syntactic schemata for both verbs differ and even the respective obj positions differ. Nonetheless, both schemata contain

a position that is a superset of the type [np(accgen),cp(int)], which allows the sentence to be accepted.

8 Discontinuous structures

Generally speaking, discontinuity is a source of problems for grammatical descriptions that are both constituency and dependency based, since it causes some tree edges to cross. Thus description of discontinuous structures can be seen as another example of an advanced feature of the grammar.

Based on the analysis of the preliminary version of Składnica (Woliński et al. 2011), we have augmented the current version of Świgra with rules for discontinuous structures which seem common in Polish sentences. This includes two main types of discontinuity.

The first type could be called inflectional. It involves syntactic forms of verbs (level 1 earlier). For example, in the following sentence the future form *będzie rosnąć* of the verb ROSNĄĆ 'to grow' is discontinuous: an adverb was inserted between its constituents:

(23) *Ale ich liczba w Europie będzie szybko rosnąć.*
but their number in Europe will quickly grow

'But their number will quickly grow in Europe.'

Similar problems concern analytic forms of the past tense, imperative mood with the particle *niech*, and conditional mood with the particle *by*.

The second type of discontinuity concerns arguments **fw** which get separated from their head as in the example:

(24) *Metody te* (methods these) *można* (is possible) *podzielić na dwie grupy* (divide into two groups).

'It's possible to divide these methods into two groups.'

The phrases *metody te* and *na dwie grupy* are dependants of the verb *podzielić*. However, the first is separated from *podzielić* by its head *można*. An important element of this pattern is that the argument phrase is separated from its head by the head of the containing phrase.

The implemented mechanism applies to arguments of infinitives (examples 24 and 25), passive participles or adjectives in the predicative position (26), and nouns (27).

(25) [*Ubogiej oświacie* poor education] *rządzący* rulers *chcieli* wanted [*zabrać* take away *200 tysiący* 200 thousands].

'The rulers wanted to
[take away 200 thousand [zlotys] from underfunded education].'

(26) *Jeszcze do niedawna proastmin* yet until recently proastmin [*dostępny* available] *był* was [*bez recepty* without prescription].

'Until recently, proastimin was [available without prescription].'

(27) [*Od zwracanego cła wyrównawczego* from refunded duty countervailing] *nie płaci się* not pay refl [*odsetek* interest].

'One does not pay [interest on refunded countervailing duty].'

The separated phrase moves to the initial (24, 25, 27) or the final (26) position in the containing phrase one level higher.

In the grammar rules, we have assumed that only one argument of the given predicate can be moved in this way (cf. Maier and Lichte 2011). Moreover, we restrict the linear order of the phrases to those shown above. If D is a dependant of R being moved within C (C is always a clause), we allow two variants. In the first, D is the initial constituent of C and R follows the head of C as in examples (24: D = *Metody te*, head of C is *można*, R = *podzielić na dwie grupy*), (25), and (27). In the second, R precedes the head of C and D is the final constituent of C—example (26): R = *dostępny*, head of C is *był*, D = *bez recepty*.

Figure 3 shows a sentence containing discontinuities of both kinds being discussed. The meaning 'to meet' can be expressed with a reflexive construction with the verb *zobaczyć*. The reflexive marker *się* is a dependant of the verb *zobaczyć*. But the form *zobaczyć* is separated from the reflexive marker by the main verb *mogli* 'be able'. To make the example more complicated, the reflexive marker is placed in the middle of an analytic form of the verb *będziemy mogli* 'will be able'. This sentence sounds very natural to the Polish ear. The strange word order, with the reflexive marker inside an analytic form of the other verb sounds even better than the variant where both phrases are continuous:

(28) *Będziemy mogli znów się zobaczyć z Piotrem w Krakowie.*
will be able again refl see with Peter in Kraków

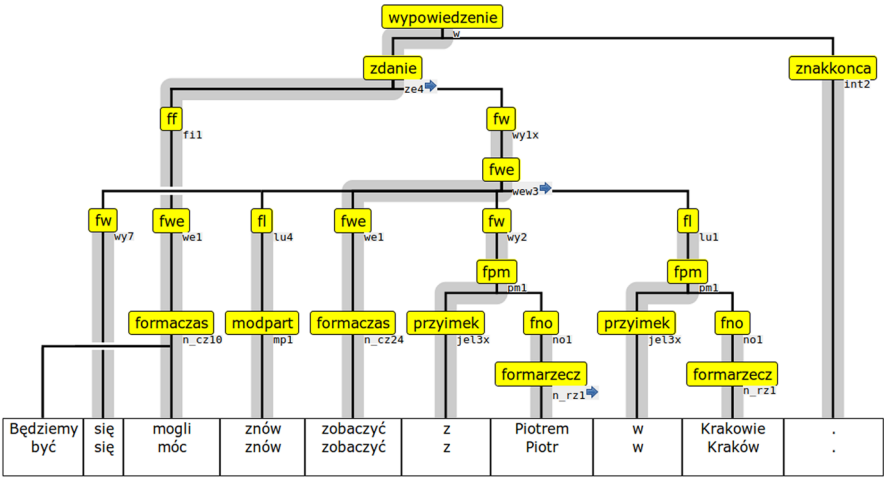


Fig. 3 A tree with crossing branches corresponding to a discontinuous structures

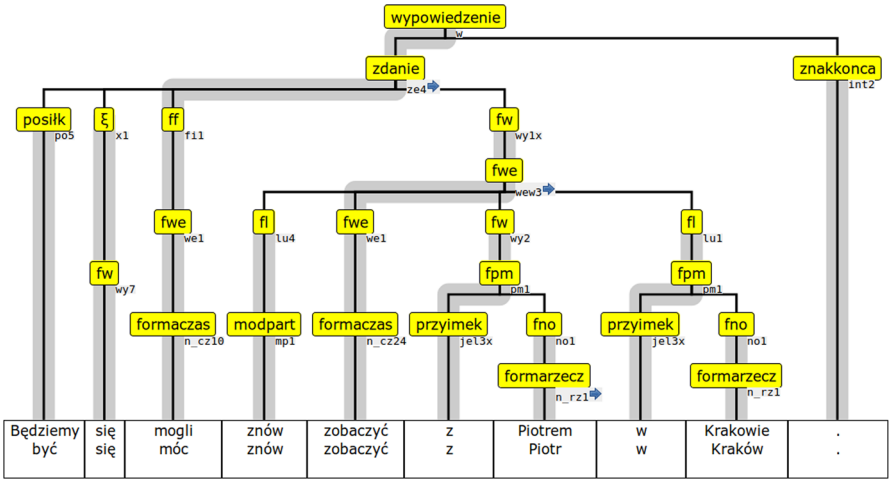


Fig. 4 The continuous structure used by Świgra to represent the sentence from Fig. 3

The ability to generate such structures is not strictly speaking facilitated by the change of the valency dictionary. However, since to deploy Walenty we needed to redesign the mechanism for filling valency slots, this was a good moment to extend this mechanism to allow some arguments to migrate up the tree.

Since the tools used to build the treebank are not well suited to discontinuous constructions, the structure of Fig. 3 is represented in a continuous form shown in

Fig. 4. Non-terminal **posiłk** is used to represent an auxiliary part of a form that can move within the clause headed by this form. In the example, it is future auxiliary *będziemy*. The representation of the second discontinuity is more complicated.

The moved argument **fw**, being the reflexive marker *się* in the example, becomes the only constituent of a special non-terminal unit labelled ξ. This phrase is an argument of the verbal phrase with the head *zobaczyć*, but to make the tree continuous it has to be moved one level higher and become a dependant of *mogli*. The unit ξ signals that this constituent is 'alien' ($\xi\acute{\epsilon}\nu o\zeta$). Information that a constituent has migrated is passed using the attributes. One of the attributes, rekcja, lists arguments realised in a given phrase. The reflexive marker is not listed as an argument of *mogli*, which has only one argument [infp(perf)]. However, it is listed as an argument of *zobaczyć*: [sie, prepnp(z,inst)]. A special value infp(perf)/[sie] is also used as the attribute tfw of the argument phrase **fw** *znów zobaczyć z Piotrem w Krakowie*. This value represents an infinitival phrase infp(perf) with a missing reflective marker sie. When parsing, Świgra checks that the element marked as ξ matches the specification of the gap in its sibling infp(perf).

It is worth noting that conversion between the trees in Figs. 3 and 4 is completely deterministic.

9 Adapting Składnica to Walenty

The core reason for using Walenty in Świgra was to introduce its rich information to the Składnica treebank. But that required some operations to be performed on the treebank.

Składnica is being developed using a system named Dendrarium, which allows trees generated with Świgra to be manually disambiguated and validated (Woliński 2010). The development is iterative and the system includes a module to automatically re-annotate a parse forest generated with a changed grammar preserving the tree previously chosen by annotators. However, in the form previously implemented, the system looked for a tree that was literally identical to the one previously selected. Because of new features in Walenty, some systematic changes had to be allowed between the old and the new trees. So, to adapt Składnica to Walenty, an algorithm was implemented that accepts the tree as matching if it differs only in a pre-specified way from the previously selected one.

To make the upgrade procedure easier to manage, the changes required for adopting Walenty were split into a few sets of independent changes, which were applied incrementally. Each set of changes was tested against the treebank and necessary corrections were performed. The corrections involved the rules of the grammar, valency schemata of Walenty, or arguments selected for particular sentences in the treebank. This way, all three resources were tested against each other.

In the first step, Walenty was mapped to a form close to the original dictionary with the intention of detecting incompatible differences in valency schemata. At this stage phrase types of Walenty were mapped back to the system used previously; all xp(...) phrases were mapped to generic advp; and lexical heads were introduced in the grammar and confronted with lexicalised schemata of Walenty. After re-parsing

of the corpus, schemata from Walenty were confronted with arguments selected by annotators.

At the beginning of procedure, there were 10,673 accepted trees in Składnica. The tree previously accepted by the annotators was found among new parses in 10,193 cases (95.5%). For the remaining 480 sentences (4.5%), the parser using Walenty did not produce a compatible tree (in 255 cases (2.4%) the new parse forest was empty). Analysis has shown that these sentences exhibit a wide range of problems including errors in both Składnica and in Walenty. For some verbs in particular, the two dictionaries differ as to whether a given dependant should be considered a complement or an adjunct. Another difference consists in modifying the original schema to its phraseological version. We have decided to upgrade the rest of the treebank and present those problematic sentences for a new assessment of treebank annotators. So, the following procedures were performed on the set of 10,193 trees.

In the following steps, which were mostly automatic, the symbols used for types of phrases were made consistent with Walenty and the subj label was added to respective phrases.

The last step was devoted to the introduction of semantically motivated xp(...) phrases. The advp specification in the old dictionary was very general: this type of phrase could be realised by any adverbial phrase or any prepositional-nominal phrase prepnp. The annotators were free to decide whether a particular prepositional phrase can be interpreted as advp in a given context. We expected many problems in matching these types.

It turned out that in about 130 sentences, some of the advp phrases in the old trees did not match any subtype of xp in the new ones. The list of sentences with this problem was analysed and the problems resolved in one of the following ways:

1. The old advp was replaced in the treebank with a specific prepositional phrase in accordance with a schema present in Walenty. For instance, the schema of the verb JECHAĆ 'ride' used in example (29) contains three xps: xp(abl), xp(adl) and xp(perl) as counterparts of two advp in the old dictionary whereas the phrase *w audi* 'in the audi' does not represent any of them and should be interpreted as a prepnp(w, loc).
2. A schema of Walenty needed to be amended by a particular subtype of xp or prepnp phrase. For example, all schemata of the verb ZWRÓCIĆ SIĘ were connected with its meaning 'talk to, ask'. However, sentence (30) contains the verb in the meaning 'turn', which requires xp(adl) realised in the sentence by the phrase *w stronę Wiktora* 'towards Wiktor'. Thus a new schema for the verb was added.
3. The offending phrase was changed from an argument **fw** to an adjunct **fl** in the treebank. For instance, the phrase *z nim* 'with him' in sentence (29) was previously interpreted as the realisation of the second advp argument of JECHAĆ 'ride', whereas it is actually an adjunct.
4. A new realisation for some subtype of xp had to be added. For example, on the basis of sentence (30), the comprepnp(w stronę) 'towards' was added as a possible realisation of the type xp(adl).

The rather extreme example (29) shows that deciding whether a particular prepnp is an argument or adjunct based on such general information as advp is really hard.

Therefore, using much more precise information as xp phrases provide would help future annotators of Składnica and minimise possible mistakes.

(29) *Zginął* *25-letni* *kolega* *kierowcy,* *który*
die.SG.PAST 25-year-old.NOM.SG colleague.NOM.SG driver.GEN.SG who.NOM.SG
jechał *z* *nim* *w* *audi.*
go,ride.SG.PAST with.INST he.INST.SG in.LOC audi.LOC.SG

'A 25-year-old colleague of the driver, who was riding with him in the audi, died.'

(30) *Ludzie* *odeszli* *i* *zwrócili* *się* *w* *stroną*
people.NOM.PL walk away.PL.PAST and turn.PL.PAST REFL in.ACC direction.ACC.SG
Wiktora.
Wiktor.GEN.SG

'People walked away and turned towards Wiktor.'

Another type of problem that showed up in the process was the ambiguity of the advp specification. Some phrases can be interpreted as xp of various subtypes. For example *gdzieś* 'somewhere' can be xp(loc)—locative or xp(adl)—ablative. The phrases *przez most* 'through a bridge', *przez godzinę* 'during one hour' and *przez niego* 'because of him' all are prepnp(przez,acc) in Polish, so they all qualify as syntactically plausible realisations of xp(perl), xp(dur), or xp(cause), but only the first is really perlative (it expresses a path of a movement), the second—durative, and the third—causative. The list of about 200 sentences containing such ambiguities was given to an expert, who decided which interpretation to choose for each of them. Real ambiguity appears if schemata of a verb contain more than one xp with the same realisation. For instance, sentence (31) contains the phrase *po południu* 'in the afternoon' being temporal realisation of prepnp(po,loc). The same phrase type belongs to realisations of locative and perlative phrases. In particular, the corresponding schema of the verb *dziać się* 'happen' contains two of them—xp(temp) and xp(locat), cf. the locative phrase *po domach* 'at home' in sentence (32). Therefore, the ambiguity between locative and temporal interpretation of the phrase had to be resolved for sentence (31).

(31) *Działo* *się* *to* *po* *południu.*
happen.SG.PAST .REFL it.NOM.SG after.LOC noon.LOC.SG

'It happened in the afternoon.'

(32) *Niejedna* *rewolucja* *dzieje* *się* *po*
more than one.NOM.SG revolution.NOM.SG happen.SG.PRES .REFL in.LOC (distributional)
domach.
home.LOC.PL

'Many revolutions happen at home.'

We are aware that some problems remain after the update procedure. Nominal phrases are not typical realisations of xp phrases. The only exception is np(inst), which is a possible realisation of xp(dur) (*czekać godzinami* 'to wait for hours') and xp(perl) (*jechać drogą* 'to drive along the road'). Such realisations were absent in the old valency dictionary, so such phrases were considered adjuncts in the treebank. These could be changed to respective xp now. Moreover, for some verbs of movement, which allow for an xp(perl) argument, the schemata of Walenty contain both xp(perl) and np(inst) (*jechać samochodem* 'to drive a car'). Only the np(inst) argument was present in the old dictionary, and could be used for both types of arguments. On the other hand, the opposite change that is described in item 3—from adjunct **fl** to argument **fw**—was not detected by this procedure, as prepnps are always admissible as an adjunct.

Unfortunately, occurrences of these problems could not be detected automatically. To make the annotation consistent with Walenty, some more manual corrections will be needed.

10 Evaluation

We begin by taking a closer look at the process mapping arguments of type advp to corresponding xps, described in Sect. 9. The columns of Table 3 correspond to two phases of this process: actions taken in cases where no corresponding xp was found in schemata of Walenty and where several xp types matched. Both phases resulted in changes applied to Składnica (the middle part of the table) or to Walenty (the lower part).

The upper part of the table shows a summary of the process. The first row contains the numbers of sentences processed in each phase, with percentages calculated w.r.t. the number of all sentences undergoing the upgrade, i.e. 10,193, cf. Sect. 9. The percentage in the following two rows is given w.r.t. these ones. The percentage in the other two parts of the table is calculated w.r.t. the total numbers of manual corrections (the numbers in bold).

As was expected, in the case of multiple matching xp types, simple disambiguation sufficed for most sentences, namely 182 (86%) cases. In the remaining sentences, in both phases, a correction was necessary in one or both resources.

In the case of corrections in Składnica, the most frequent decision was to replace an advp with a specific prepnp. A special case of such a replacement is the verb *być* 'to be' and its iterative counterpart *bywać*, which accept any prepositional phrase as its argument. In both phases of the process, these two types of changes were made in 58 sentences – 17% in total or 36,5% of "hard" cases. For 11 sentences (7%) advp argument was reinterpreted as an adjunct, 4 sentences (2.5%) involve changing advp to a lexicalised argument.

Please note that for corrections in Walenty the table includes the numbers of sentences that triggered a change, not the number of changes. For instance, as many as 9 sentences were 'cured' by adding comprepnp(w kierunku) 'in direction' as a possible realisation to the type xp(adl). On the other hand, a problem reported in one sentence

Table 3 Details of the process of mapping the old type advp to xp in Składnica

	No matching xp		Several matching xps	
	Number	%	Number	%
Considered sentences	130	1.28	211	2.07
Plain disambiguation	–		182	86.26
Correction needed	**130**	100.00	**29**	13.74
Resolved by correcting the tree in Składnica				
Phrase type changed to prepnp	26	20.00	2	6.90
Accepted as generic prepnp for *być* 'to be'	16	12.31	14	48.28
Replaced with a lexicalised type	4	3.08	0	0.00
Changed to other argument type	4	3.08	0	0.00
Argument replaced with an adjunct	9	6.92	2	6.90
Lack of parse	8	6.15	1	3.45
Other incompatibility	12	9.23	4	13.79
Corrections total	79	60.77	23	79.31
Resolved by correcting Walenty				
Changes in schemata				
New schema with xp	10	7.69	1	3.45
Schema augmented with xp	11	8.46	2	6.90
New schema with prepnp	3	2.31	0	0.00
Schema augmented with prepnp	1	0.77	0	0.00
New idiomatic schema	7	5.38	0	0.00
Added realisations of some xp				
New comprepnp	13	10.00	3	10.34
New prepnp or advp	7	5.38	5	17.24
Corrections total	52	40.00	11	37.93

often resulted in a cascade of changes in related entries of Walenty (aspectual pairs, synonyms etc.) to keep the dictionary's integrity.

To sum up, 102 sentences (79+23, 30% of of 341 "problematic" sentences) required a manual change of annotation in Dendrarium. The remaining 239 sentences (70%) were mapped semiautomatically, after disambiguating xp and reparsing Składnica with the corrected version of Walenty.

The present version of Składnica contains human-validated trees for 11,938 sentences consisting of 131,334 tokens (including punctuation). Table 4 shows the numbers of sentences accepted by various versions of Świgra. As can be seen, the parser over-generates and not all of the generated trees get subsequently accepted by annotators in manual validation. As a result of the switch to Walenty, the parser accepted 909 more sentences, while the gain in validated trees is 1265. This proves that the new dictionary allowed some analyses previously rejected by annotators to be corrected. In other words, the tendency of the parser to over-generate is lower in the new version. The difference between structures accepted by the parser and by humans

Table 4 Coverage of various versions of Świgra counted on the 20,000 sentence corpus of Składnica

	Accepted by parser		Positively validated by annotators			
	Sentences	%	Sentences	%	Tokens	Avg sent. length
Świdziński's grammar		$\simeq 30$				
Świgra 2 before Walenty	13,194	66.0	10,673	53.4	112,297	10.49
Świgra 2 with Walenty	14,103	70.5	11,938	59.7	131,334	11.00

Table 5 Use of valency schemata by type of predicate

	Sentences	Phrases
All sentences	11,938	
Valency-sensitive phrases	11,481	42,645
Non-empty	10,708	41,516
Verbal and de-verbal	11,095	20,701
Non-empty	10,706	19,039
Nominal	8371	21,021
Non-empty	271	346
Adjectival	870	923
Non-empty	67	71
xp arguments	1828	2005
Multiple realised xp arguments	30	30

is smaller for the Walenty version by about 14% of cases. We can guess that the parser missed valency schemata for about that amount of sentences, which caused it, e.g., to classify some arguments as adjuncts. The newly accepted sentences include those with verbs missing from the old dictionary, but some new successes are also due to new features of the parser discussed in Sects. 6–8.

For the present version of the parser almost 60% of sentences get validated interpretations. This number may seem low, but it is worth noting that Składnica is the only treebank of Polish for which the completeness of underlying language description can be assessed. The LFG treebank mentioned in Sect. 1 contains only a subset of Składnica sentences that POLFIE was able to parse plus some parsable sentences drawn from a much larger corpus. Such construction of the treebank provides no clue on the coverage of the grammar with respect to a fixed corpus. A similar approach was taken in the case of dependency treebanks: converted validated trees of Składnica 0.5 were amended with trees for interesting sentences selected by hand. This may introduce a bias in what the dependency parser learns, since some constructions can be systematically missing from the hand picked sentences.

In the remaining part of this section we are going to investigate how schemata of Walenty 'work' in Składnica. Table 5 shows counts of all phrases in Składnica in which valency frames are used. The 'non-empty' rows show how many times a non-empty subset of the schema is used, i.e. how often does a given type of predicate take at least one argument. As can be seen, verbs usually take dictionary-determined arguments—only 1662 of 20,701 verbal predicates (8%) have no dependants or only adjuncts. On

Table 6 Corpus frequency of various argument types of adjectives, nouns and verbs in Składnica

Types of arguments	Predicates			
	Verbs	Nouns	Adjectives	Total
np	18,560	161	33	18,754
prepnp	2747	49	13	2809
prepncp	30	7	0	37
cp	1137	106	19	1262
ncp	4	1	0	5
sie	2356	–	–	2356
infp	1854	–	4	1858
xp	2024	10	2	2036
advp	114	–	0	114
adjp	1191	0	0	1191
prepadjp	20	0	0	20
or	430	18	0	448
comprepnp	3	4	0	7

the other hand, only 346 of 21,021 nominal heads of phrases in Składnica (1.6%) use a non-empty schema from Walenty. The rest of the nominal phrases contain only typical dependants (adjectives and nouns in the genitive), which can be considered adjuncts. For adjectives, the number is 7.7%. Thus, decent description of verbs is the most important feature of a valency dictionary for Polish. However, parsing would fail for about 338 (271+67) sentences of Składnica without non-verbal schemata in Walenty.

The last two rows of Table 5 show how often xp phrases introduced by Walenty are used in Składnica—in about 5% of non-empty schemata used. The xp subtypes allow differentiation among various 'adverbial' arguments, so we were interested how often more than one xp is realised for a given predicate. As it turns out, there are only 30 such phrases in Składnica.

The number of various types of arguments required by verbs, nouns and adjectives is summarised in Table 6. As one might expect, the most frequent type of argument is nominal phrase and prepositional-nominal phrase (but almost 7 times less often). Observe that the proportion between types of arguments is similar for all types of predicates. Reflexive marker *się* does not appear with nouns and adjectives (we consider gerunds and participles as verbal forms for this table). Similarly, Polish nouns do not have infinitival and adverbial arguments. Furthermore, we have checked that adverbial, adjectival, and prepositional adjectival arguments, which are absent in Składnica for nouns and adjectives, are rare in Walenty as well. Obviously, it was very unlikely we would find them in Składnica.

In Table 7 we analyse the frequency of various subtypes of xp. The phrases describing location of an action turned out to be most frequent. The adl, abl, perl triple is often used with verbs of movement. And it seems that specifying destination is most important for speakers, while the trajectory of movement is specified least often. Another

Table 7 Types of xp phrases occurring in Składnica

	No. of phrases
xp(locat)—location	823
xp(adl)—'to' point of a movement	730
xp(abl)—'from' point of a movement	206
xp(perl)—trajectory of a movement	46
xp(mod)—manner	139
xp(temp)—time	57
xp(dur)—duration	14
xp(dest)—aim	14
xp(caus)—cause	4
xp(instr)—instrument	3

Table 8 Types of subject phrases occurring in Składnica

	No. of phrases
np(nom)—standard nominal	9153
infp—infinitival phrase	52
cp(że)—clause with *że* 'that'	75
Other clauses	8

relatively frequent type is xp(mod) specifying the manner of performing some action. The number for time-related features is lower, but this is because the time of an action is usually expressed with an adjunct. Walenty only uses xp(temp) and xp(dur) with verbs such as 'to begin', 'to end', 'to last'.

Another feature that differentiates Walenty from the previously used dictionary is non-nominal subjects. Table 8 shows, how often this feature is used in Składnica. As it turns out, the cp(że) type illustrated by sentence (14) on page 12 is the most common type of non-nominal subject. On the other hand, non-nominal subjects constitute only 1.5% of all subjects (this number can be slightly biased by the late adoption of the concept).

11 Conclusions and perspectives

Składnica is the first constituency treebank of Polish of a considerable size. The resource is now coupled with an independently developed valency dictionary, which marks an important turning point in its development. The fact that Walenty is actively maintained makes further development of the parser easier. From the other point of view, Składnica provides verification for schemata of Walenty.

The current version of Składnica can be downloaded from the address http://zil.ipipan.waw.pl/Sk%C5%82adnica. The treebank is available as a set of 20,000 XML files with a simple ad-hoc DTD. Each file contains the complete parse forest generated by Świgra 2 (empty if the parser failed to recognise the sentence). If a given sentence

was accepted by the annotators, the fact is marked in meta-data and the correct tree is marked in the forest. The present version of the treebank is also available for easy access in the treebank search engine: http://treebank.nlp.ipipan.waw.pl/.

The new version of Składnica will also be converted to the dependency form and used for training dependency parsers. An interesting question is whether the new features of the treebank (in particular types of xp phrases) can help in training statistical disambiguation tools and parsers. Another direction of development is to use the semantic layer of Walenty to generate predicate-argument structures using semantic role labels.

Acknowledgements Work partly financed as part of the investment in the CLARIN-PL research infrastructure funded by the Polish Ministry of Science and Higher Education. The development of previous versions of Składnica was funded by the Polish National Science Centre.

References

Böhmová, A., Hajičová, E., Hajič, J., & Hladká, B. (2003). The Prague dependency treebank: A three-level annotation scenario. In A. Abeillé (Ed.), *Treebanks: Building and using parsed corpora, language and speech* (pp. 103–127). Dordrecht: Kluwer Academic Publishers.

Fillmore, C. J., Johnson, C. R., & Petruck, M. R. L. (2003). Background to FrameNet. *International Journal of Lexicography*, *16*(3), 235–250.

Hajič, J. (2005). Complex corpus annotation: The Prague dependency treebank. In M. Šimková (Ed.), *Insight into Slovak and Czech Corpus Linguistics* (pp. 54–73). Bratislava: Veda.

Hajnicz, E., Andrzejczuk, A., & Bartosiak, T. (2016a). Semantic layer of the valence dictionary of Polish Walenty. In N. Calzolari, K. Choukri, T. Declerck, M. Grobelnik, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation, LREC 2016, ELRA* (pp. 2625–2632). Portorož: European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2016/index.html.

Hajnicz, E., Patejuk, A., Przepiórkowski, A., & Woliński, M. (2016b). Walenty: słownik walencyjny j̨ezyka polskiego z bogatym komponentem frazeologicznym. In K. Skwarska & E. Kaczmarska (Eds.), *Výzkum slovesné valence ve slovanských zemích* (pp. 71–102). Prague: Slovanský ústav AV ČR.

Kettnerová, V., Lopatková, M., & Bejček, E. (2012). The syntax-semantics interface of Czech verbs in the valency lexicon. In *Proceedings of the 15th EURALEX International Congress* (pp. 434–443). Oslo: Department of Linguistics and Scandinavian Studies, University of Oslo.

Kingsbury, P., & Palmer, M. (2002). From TreeBank to PropBank. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002)* (pp. 1989–1993). Las Palmas, Spain

Kipper, K., Korhonen, A., Ryant, N., & Palmer, M. (2008). A large-scale classification of English verbs. *Languge Resources and Evaluation Journal*, *42*(1), 21–40.

Lewandowska-Tomaszczyk, B., Górski, R., Łaziński, M., & Przepiórkowski, A. (2013). The National Corpus of Polish (NKJP). Language use and data analysis. In I. KorChahine & C. Zaremba (Eds.), *Travaux de slavistique : Actes du VIe congrès de la Slavic Linguistic Society* (pp. 309–319). Aix-en-Provence: Presses Universitaires de Provence.

Maier, W., & Lichte, T. (2011). Characterizing discontinuity in constituent treebanks. In P. Groote, M. Egg, & L. Kallmeyer (Eds.), *Formal Grammar: 14th International Conference, FG 2009, Bordeaux,*

France, July 25–26, 2009, Revised Selected Papers (pp. 167–182). Berlin: Springer. https://doi.org/10.1007/978-3-642-20169-1_11.

Marcus, M. P., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, *19*(2), 313–330.

Palmer, M., Kingsbury, P., & Gildea, D. J. (2005). The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, *31*(1), 71–106.

Patejuk, A., & Przepiórkowski, A. (2014). Synergistic development of grammatical resources: A valence dictionary, an LFG grammar, and an LFG structure bank for Polish. In V. Henrich, E. Hinrichs, D. de Kok, P. Osenova, & A. Przepiórkowski (Eds.), *Proceedings of the Thirteenth International Workshop on Treebanks and Linguistic Theories (TLT 13)* (pp. 113–126). Department of Linguistics (SfS), University of Tübingen, Tübingen. http://tlt13.sfs.uni-tuebingen.de/tlt13-proceedings.pdf.

Pereira, F., & Warren, D. H. D. (1980). Definite clause grammars for language analysis—A survey of the formalism and a comparison with augmented transition networks. *Artificial Intelligence*, *13*, 231–278.

Przepiórkowski, A. (2004). O wartości przypadka podmiotów liczebnikowych. *Biuletyn Polskiego Towarzystwa Językoznawczego*, *LX*, 133–143.

Przepiórkowski, A., & Patejuk, A. (2020). From Lexical Functional Grammar to enhanced Universal Dependencies: The UD-LFG treebank of Polish. *Language Resources and Evaluation*, *54*, 185–221. https://doi.org/10.1007/s10579-018-9433-z.

Przepiórkowski, A., Bańko, M., Górski, R. L., & Lewandowska-Tomaszczyk, B. (Eds.). (2012). *Narodowy Korpus Języka Polskiego*. Warsaw: Wydawnictwo Naukowe PWN.

Przepiórkowski, A., Hajnicz, E., Patejuk, A., & Woliński, M. (2014a). Extended phraseological information in a valence dictionary for NLP applications. In *Proceedings of the Workshop on Lexical and Grammatical Resources for Language Processing (LG-LP 2014)* (pp. 83–91). Dublin: Association for Computational Linguistics and Dublin City University; http://www.aclweb.org/anthology/siglex.html#2014_0.

Przepiórkowski, A., Hajnicz, E., Patejuk, A., Woliński, M., Skwarski, F., & Świdziński, M. (2014b). Walenty: Towards a comprehensive valence dictionary of Polish. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014, ELRA, Reykjavík, Iceland* (pp. 2785–2792). http://www.lrec-conf.org/proceedings/lrec2014/index.html.

Przepiórkowski, A., Skwarski, F., Hajnicz, E., Patejuk, A., Świdziński, M., & Woliński, M. (2014c). Modelowanie własności składniowych czasowników w nowym słowniku walencyjnym języka polskiego. *Polonica*, *XXXIII*, 159–178.

Przepiórkowski, A., Hajič, J., Hajnicz, E., & Urešová, Z. (2017). Phraseology in two Slavic valency dictionaries: Limitations and perspectives. *International Journal of Lexicography*, *30*(1), 1–38.

Seddah, D., Tsarfaty, R., Kübler, S., Candito, M., Choi, J. D., Farkas, R., et al. (2013). Overview of the SPMRL 2013 shared task: A cross-framework evaluation of parsing morphologically rich languages. In *Proceedings of the Fourth Workshop on Statistical Parsing of Morphologically-Rich Languages* (pp. 146–182). Seattle: Association for Computational Linguistics.

Sgall, P., Hajičová, E., & Panevová, J. (1986). *The meaning of the sentence in its semantic and pragmatic aspects*. Dordrecht: D. Reidel.

Świdziński, M. (1992). *Gramatyka formalna języka polskiego*. Rozprawy Uniwersytetu Warszawskiego, Wydawnictwa Uniwersytetu Warszawskiego, Warszawa

Świdziński, M. (1994). Syntactic dictionary of Polish verbs, manuscript, Uniwersytet Warszawski and Universiteit van Amsterdam.

Świdziński, M., & Woliński, M. (2010). Towards a bank of constituent parse trees for Polish. In P. Sojka, A. Horák, I. Kopeček, & K. Pala (Eds.), *Text, Speech and Dialogue: 13th International Conference, TSD 2010, Brno, Czech Republic* (pp. 197–204). Heidelberg: Springer-Verlag. no. 6231 in Lecture Notes in Artificial Intelligence.

Urešová, Z. (2009). Building the PDT-Vallex valency lexicon. In *Proceedings of the 5th Corpus Linguistics Conference, University of Liverpool*.

Woliński, M. (2004). Komputerowa weryfikacja gramatyki Świdzińskiego. Ph.D. dissertation, Institute of Computer Science, Polish Academy of Sciences, Warsaw.

Woliński, M. (2010). Dendrarium - an open source tool for treebank building. In M. A. Kłopotek, M. Marciniak, A. Mykowiecka, W. Penczek, & S. T. Wierzchoń (Eds.), *Proceedings of IIS'2010, Wydawnictwo Akademii Podlaskiej* (pp. 193–204).

Woliński, M. (2015). Deploying the new valency dictionary Walenty in a DCG parser of Polish. In M. Dickinson, E. Hinrichs, A. Patejuk, & A. Przepiórkowski (Eds.), *Proceedings of the Fourteenth International Workshop on Treebanks and Linguistic Theories (TLT 14), Institute of Computer Science, Polish Academy of Sciences, Warsaw* (pp. 221–229). http://tlt14.ipipan.waw.pl/proceedings/.

Woliński, M. (2019). *Automatyczna analiza składnikowa języka polskiego*. Warszawa: Warsaw University Press.

Woliński, M., Głowińska, K., & Świdziński, M. (2011). A preliminary version of Składnica—a treebank of Polish. In Z. Vetulani (Ed.) *Proceedings of the 5th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics* (pp. 299–303). Poland: Poznań.

Wróblewska, A. (2014). Polish dependency parser trained on an automatically induced dependency bank. Ph.D. dissertation, Institute of Computer Science, Polish Academy of Sciences, Warsaw.

Wróblewska, A. (2018) Extended and enhanced Polish dependency bank in Universal Dependencies format. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)* (pp. 173–182). Brussels: Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-6020, https://www.aclweb.org/anthology/W18-6020.

Wróblewska, A., & Woliński, M. (2012). Preliminary experiments in Polish dependency parsing. In P. Bouvry, M. A. Kłopotek, F. Leprevost, M. Marciniak, A. Mykowiecka, & H. Rybiński (Eds.), *Security and Intelligent Information Systems: International Joint Conference, SIIS 2011, Warsaw, Poland, June 13-14, 2011, Revised Selected Papers, Springer-Verlag, no. 7053 in Lecture Notes in Computer Science* (pp. 279–292). http://www.springer.com/computer/communication+networks/book/978-3-642-25260-0.

Žabokrtský, Z., & Lopatková, M. (2007). Valency information in VALLEX 2.0: Logical structure of the lexicon. *The Prague Bulletin of Mathematical Linguistics*, *87*, 41–60.