# The structure of a dictionary entry and grammatical properties of multi-word units

**Monika Czerepowicka**

University of Warmia and Mazury in Olsztyn (Poland)
E-mail: monika.czerepowicka@uwm.edu.pl

**Abstract**

Users of highly inflectional languages expect dictionaries to provide clear inflectional information so that the creation or use of a given form does not generate additional problems. The development of technologies and tools for machine language processing has naturally made contemporary inflectional dictionaries advanced electronic works that contain tools for the individualisation of their content in line with users' needs. The main concern of this article is the influence of the grammatical properties of language units on lexicographic description, in particular the structure of a dictionary entry. This issue will be discussed with reference to *Verbel. The Inflectional Dictionary of Polish Verbal Phrases*, which is an electronic dictionary listing over 5,000 multi-word units, giving all their paradigmatic forms directly. Although it is a specialist study providing a formal description of units, thanks to the proper structure of entries it is possible to be used also by non-specialists. The opportunity of choosing the scope of lexicographic information in the *Verbel* dictionary is guaranteed by a two-stage scheme of the entry which consists of a general and detailed description of units.

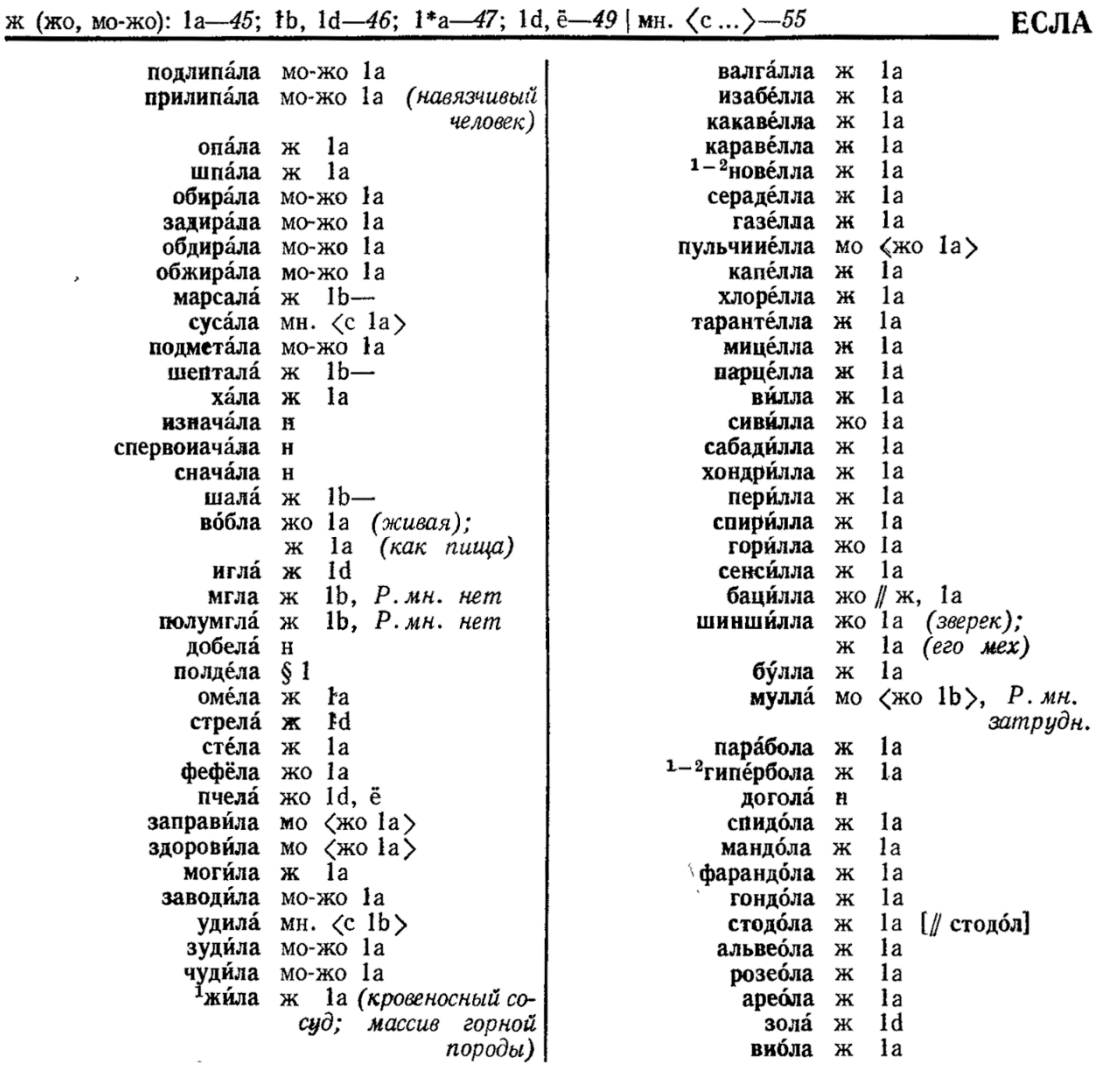
**Keywords:** multi-word units; inflection; dictionary; e-lexicography

## 1. Introduction

The subject under scrutiny is the part of lexicographic description which reports on the grammatical, mainly inflectional, information about a unit. It is assumed that language units are differentiated on the basis of their semantic and grammatical features, thus they can also be discontinuous (cf. Baldwin and Kim, 2010; Bogusławski, 1976; Mel'čuk, 2006; Mel'čuk & Zholkovsky, 1984; Sag et al., 2002). Regardless of their formal structure, however, they should be uniformly described. The position advocated in this study is that multi-word units of language should be accompanied by an equally detailed, precise and consistent inflectional description as lexemes. For this reason, a rigorous, algorithmic model will be applied to provide such a description in *Verbel. The Inflectional Dictionary of Polish Verbal Phrases* (Kosek et al., 2020).

The idea of grammatical dictionaries providing all paradigmatic forms of a unit is particularly important and useful for inflected languages, such as Polish and other Slavic languages. One can find a few methodological models that make it possible to describe units of language in an adequately detailed and precise manner and have been used in dictionaries. *The Grammatical Dictionary of Russian* (Rus. *Грамматический словарь русского языка*) by Andriey Zaliznyak (1977) is one of the first dictionaries of

this type. Zaliznyak's approach was highly innovative in dispensing with the construct of the morpheme and putting the notion of the paradigm in the spotlight, heralding the rise of 'word-and-paradigm' and other realisational theories in morphology (Iosad and others 2018: 176). Zaliznyak's morphological model consists in constructing paradigm forms from an abstract lexeme representation using rewrite rules. The dictionary contains about 100,000 units of language with their grammatical characteristics presented by symbols and listed in *a tergo* order (Fig. 1).



ж (жо, мо-жо): 1a—*45*; 1b, 1d—*46*; 1*a—*47*; 1d, ё—*49* | мн. ⟨с ...⟩—*55* **ЕСЛА**

**подлипа́ла** мо-жо 1a
**прилипа́ла** мо-жо 1a *(навязчивый человек)*
**опа́ла** ж 1a
**шпа́ла** ж 1a
**обира́ла** мо-жо 1b
**задира́ла** мо-жо 1a
**обдира́ла** мо-жо 1a
**обжира́ла** мо-жо 1a
**марсала́** ж 1b—
**суса́ла** мн. ⟨с 1a⟩
**подмета́ла** мо-жо 1b
**шепталá** ж 1b—
**ха́ла** ж 1a
**изнача́ла** н
**спервоиача́ла** н
**снача́ла** н
**шала́** ж 1b—
**во́бла** жо 1a *(живая)*; ж 1a *(как пища)*
**игла́** ж 1d
**мгла** ж 1b, *Р. мн. нет*
**полумгла́** ж 1b, *Р. мн. нет*
**добела́** н
**полде́ла** § 1
**оме́ла** ж 1b
**стрела́** ж 1d
**сте́ла** ж 1a
**фефёла** жо 1a
**пчела́** жо 1d, ё
**заправи́ла** мо ⟨жо 1a⟩
**здорови́ла** мо ⟨жо 1a⟩
**моги́ла** ж 1a
**заводи́ла** мо-жо 1a
**удила́** мн. ⟨с 1b⟩
**зуди́ла** мо-жо 1a
**чуди́ла** мо-жо 1a
**[1]жи́ла** ж 1a *(кровеносный сосуд; массив горной породы)*

**валга́лла** ж 1a
**изабе́лла** ж 1a
**какаве́лла** ж 1a
**караве́лла** ж 1a
**[1–2]нове́лла** ж 1a
**сераде́лла** ж 1a
**газе́лла** ж 1a
**пульчиие́лла** мо ⟨жо 1a⟩
**капе́лла** ж 1a
**хлоре́лла** ж 1a
**тарантéлла** ж 1a
**мице́лла** ж 1a
**парце́лла** ж 1a
**ви́лла** ж 1a
**сиви́лла** жо 1a
**сабади́лла** ж 1a
**хондри́лла** ж 1a
**пери́лла** ж 1a
**спири́лла** ж 1a
**гори́лла** жо 1a
**сенси́лла** ж 1a
**баци́лла** жо // ж, 1a
**шинши́лла** жо 1a *(зверек)*; ж 1a *(его мех)*
**бу́лла** ж 1a
**мулла́** мо ⟨жо 1b⟩, *Р. мн. затрудн.*
**пара́бола** ж 1a
**[1–2]гипе́рбола** ж 1a
**догола́** н
**спидо́ла** ж 1a
**мандо́ла** ж 1a
**фарандо́ла** ж 1a
**гондо́ла** ж 1a
**стодо́ла** ж 1a [// стодо́л]
**альвео́ла** ж 1a
**розео́ла** ж 1a
**арео́ла** ж 1a
**зола́** ж 1d
**вио́ла** ж 1a

Figure 1. Entries from *The Grammatical Dictionary of Russian* (Rus. *Грамматический словарь русского языка*) by A. Zaliznyak (1977)

Paradigms can also be shown precisely by illustrative tables. This way of data presentation is used in grammatical dictionaries of verbal units of French (Bescherelle, 1978) or Polish (Saloni, 2007). All verbs are arranged in groups distinguished on the basis of their morphological structure and inflectional properties. A total of 106 patterns were identified following in-depth and detailed analyses of the Polish conjugation. The paradigm of each pattern is presented with an example verb in a

table. Thanks to proper presentation of the formal structure of a given verb and detailed instructions of recognising the morphological verb pattern the user can inflect other verbs belonging to the same group even not noted in the dictionary (Fig. 2).

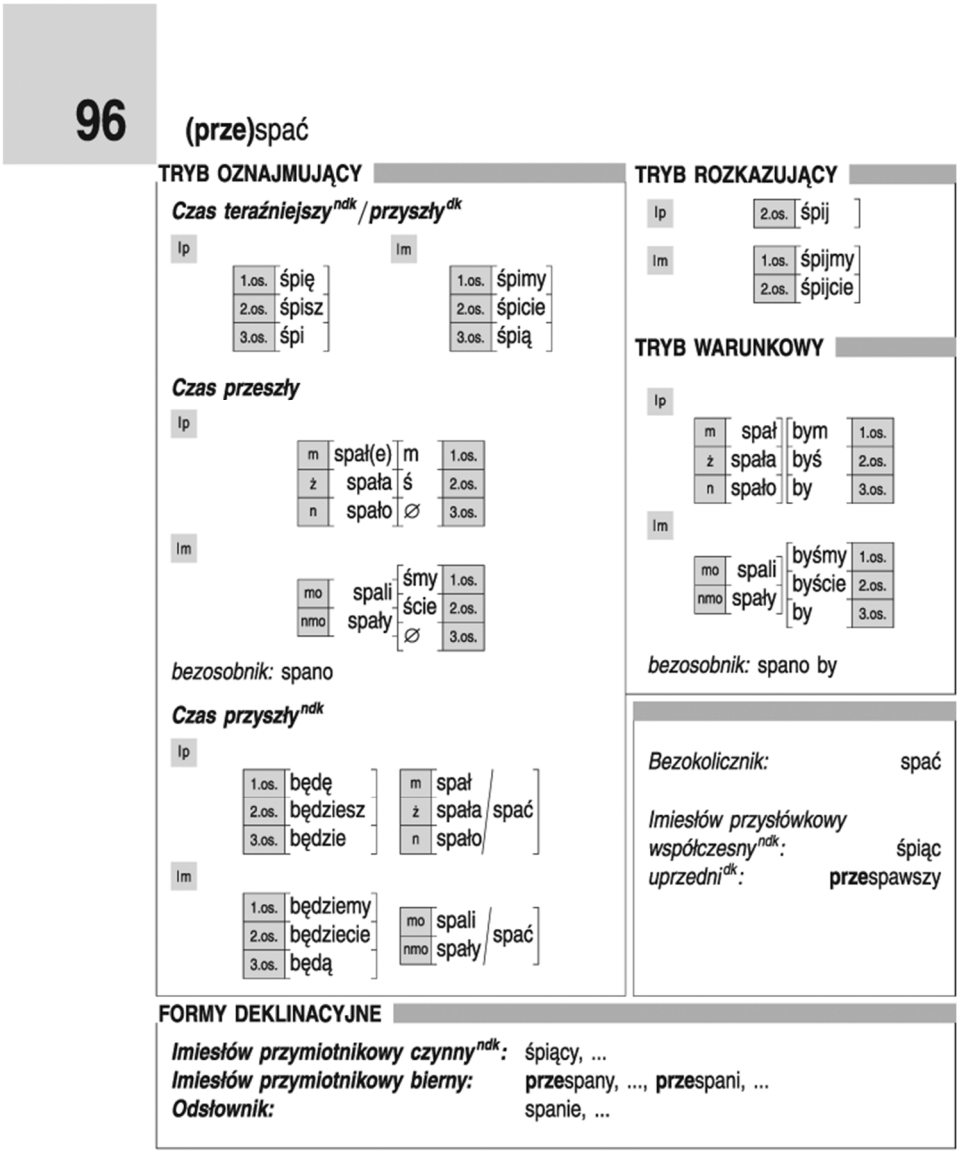



Figure 2. A table from *The Polish Verb. Inflection, Dictionary of 12,000 lexemes* (Pl. *Czasownik polski. Odmiana, słownik 12,000 czasowników*) of Z. Saloni (2001)

The development of electronic dictionaries gives an obvious opportunity to note the paradigms of all units directly (*in extenso*). However, the lexicographic description should present the nature of each unit in all its inflectional complexity. Tools to construct an appropriately precise scheme to provide the inflectional information of Polish units are included in the concept of a morphological description, proposed by Janusz S. Bień and Zygmunt Saloni (1982). The methodological perspective adopted by the authors proved to be effective in the case of lexemes, which was confirmed by *The Grammatical Dictionary of Polish* (Pl. *SGJP*; Saloni et al., 2015), which contains

descriptions of over 300,000 Polish units. This theoretical model has proven successful in machine processing as well, being used in the morphosyntactic marking of the National Corpus of Polish (Pl. NKJP; Przepiórkowski et al., 2012). However, it should be emphasised that it concerns lexemes. Since multi-word units require an equally detailed and rigorous description, as mentioned above, it has been decided to implement the model in the inflectional dictionary of Polish verbal phrases.

In this paper, terms such as "phrase", "phraseologism" and "multi-word units" are applied to refer to discontinuous units of language. Verbal units of this type can be defined as connections of at least two words that perform the function of the centre of the sentence, similarly to verbal lexemes. Because of the degree of unification of unit components, the possibility of replacing some of them and resultant changes of meaning, one can distinguish idioms, light verb constructions, and collocations among them. However, in this study we do not consider differences between the mentioned semantic types of verbal multi-word units but focus on inflected and morphosyntactic features and their influence on the structure of a dictionary entry. Still, we discuss morphological types of Polish multi-word units as well as the basis of the theoretical model used in the *Verbel* dictionary. The key terms it comprises are *a morphological word*, *a paradigmatic word*, *a flexeme* and *a vocabula*, and they reflect the multi-step procedure of a comprehensive grammatical description of language units.

## 2. Description model

A *morphological word* is defined here as a sequence of letters (graphemic shape; *signifiant*) interpreted grammatically and semantically (*signifié*). It is a complete linguistic sign. Its grammatical properties are determined on the basis of morphological features of each type of word – nouns, verbs, adjectives etc. Apart from traditional morphological categories, such as case, number, gender, and person, the register of morphological words also includes non-traditional categories, resulting from detailed inflectional description. These include, *inter alia*, such categories as agglutination and vocalism, both connected to each other and with inflection by person. Agglutination is a grammatical feature noted in the past tense inflection. The person-number morpheme (of the 1st or 2nd person) usually appears immediately after a verbal stem, forming one word in textual form: *robił-**em*** ('I did' masc.), *robił-**eś*** ('you did' masc.), *robili-**śmy*** ('we did' masculine personal), *robili-**ście*** ('you did' pl. masculine personal). Still, these morphemes can be torn off the verb stem and glued to another word in a sentence, e.g.: *Blat miał**eś*** - *Blat**eś*** *miał.* ('You had a tabletop'). There are syntactic constructions where this kind of operation is required, such as in some dependency sentence phrases: *Jan chciał, żeby**śmy poszli** na spacer.* - **Jan chciał, żeby **poszliśmy** na spacer* ('Jan wanted us to go for a walk').

The vocalism category is also observed in the past tense. The shape of the agglutination morpheme depends on the ending of the verb stem. If the stem ends in a consonant, like in masculine forms, e.g. *robił* ('he did'), the agglutination morpheme becomes vocal:

*robił**em*** (‘I did’ masc; the first-person is created by adding the agglutination morpheme to the past verb stem). If the verb stem ends in a vowel (as it is in non-masculine forms), the agglutination morpheme becomes non-vocal, e.g. *robiła**m*** (‘I did’ fem.), *robiły**śmy*** (‘we did’ non-masculine). When generating Polish verb forms, all such subtle morphological features must be taken into account.

Polish verbal morphological words are heterogeneous. They cannot be classified by the same morphological categories. Apart from formal signs, they differ in semantic identification. The full paradigm of the verb includes verbal adjectives (participles) and nouns, i.e. forms that are inflected by cases, in addition to forms inflected by person and number. Furthermore, conjugation forms are subject to morphological categories to varying degrees – for example, the category of genus is manifested in the past tense (*robił* ‘he did’, *robiła* ‘she did’), the conditional mood (*zrobiłbym* ‘I would do’ masc., *zrobiłabym* ‘I would do’ fem.), and certain complex future forms (*będę robił* ‘I will be doing’ masc., *będę robiła* ‘I will be doing’ fem.). In the case of other verb forms, gender neutralisation can be noted. Among verbal paradigmatic forms there are also those that cannot be assigned any other grammatical category than aspect, these are: infinitive, adverbial participles, impersonal forms (Pl. *bezosobnik*, forms with *-no*, *-to*), e.g. *robić* ‘to do’, *robiąc* ‘doing’, *robiono* ‘it was done.’ This prompts us to classify them into groups that fall under the same morphological categories. We call sets of forms differentiated on the basis of the same morphological categories *flexemes* or *paradigmatic words*.

The level of complexity of the Polish conjugation system calls for a special treatment. The paradigm of verbal units consists of various types of paradigmatic words (non-past and past forms, participles, conditional forms, imperative, etc.), which form separate sub-paradigms. For example, flexemes of the past tense are inflected by person, number, gender, while non-past forms (present and future simple tense) and imperative – by person and number. The theoretical problem related to consistent morphosyntactic description of verbal units is closely related to the number of flexemes belonging to a given unit, and thus to the multitude and variety of inflectional forms. In the case of verbal units, it is a systemic phenomenon, which ultimately determines the architecture of a dictionary entry which becomes a super-class – a *vocabula*. A vocabula, i.e. a dictionary entry, groups paradigmatic words with the same semantic root, so it consists of various types of flexemes characterised by different morphological features. Verbs with regular inflection patterns (full paradigm) encompass 8 or 10 flexemes (depending on the aspect).

This multistage procedure provides the basis for both the description of abstract language units and their textual realisations, at the same time providing tools for separate levels of linguistic description. This type of research perspective seems to be particularly helpful in machine language processing.

## 3. The specificity of verbal multi-word units

It should be noted that the paradigm of verbal multi-word units depends on their morphosyntactic properties as well as morphological structure. They differ in both internal syntax (mutual relations of multi-word unit components) and external syntax (matching and requirements with regard to other sentence elements; cf. Lewicki, 1986). Based on their morphosyntactic features, one can distinguish three types of Polish multi-word units. Thus, there are phrases (type 1) which are characterised by an open position for the subject in the nominative, {*ktoś*} *zbija bąki* (lit. {*someone*$_{\text{Nom}}$} *is shooting herons*, 'someone is getting lazy'), {*ktoś*} *przypina komuś łatkę*, (lit. {*someone*$_{\text{Nom}}$} *is sticking a patch on someone else*, 'someone is attributing a negative feature or behaviour to someone else'). In contrast, other phrases do not open up a position for the nominative argument (type 2). This position is permanently filled in lexically, for example, *oczy*$_{\text{Nom}}$ *wychodzą* {*komuś*} *na wierzch* (lit. {*someone's*} *eyes*$_{\text{Nom}}$ *go to the surface*, 'someone is really surprised'), *włos*$_{\text{Nom}}$ {*komuś*} *z głowy nie spadnie* (lit. *not a hair*$_{\text{Nom}}$ *will fall off* {*someone's*} *head*, 'someone will be safe'). The third group of phrases does not show any collocability with the nominative argument, which is why it is characterised by a very limited paradigm, e.g.: {*komuś*$_{\text{Dat}}$} *pada na mózg* (lit. *it falls on* {*someone's*$_{\text{Dat}}$} *brain*, 'someone acts irrationally'), {*komuś*$_{\text{Dat}}$} *przybywa na wadze* (lit. {*someone*$_{\text{Dat}}$} *has more on the scales*, 'someone is putting on weight'), {*komuś*$_{\text{Dat}}$} *brak piątej klepki* (lit. *someone*$_{\text{Dat}}$ *lacks the fifth plank*, 'someone is crazy' ). A unit's belonging to a given type determines its inflectional paradigm. Vocabulas of the first type can potentially have a full inflectional paradigm, with any limitations resulting from their semantic features (meaning). The second and third type units show numerous limitations in terms of variation by categories of person and number.

From an essentially morphological point of view verbal multi-word units can be divided into two groups: verbs and predicates. Both types differ in their formal structure and the scope of inflectional forms. The VERB class includes mainly phrases based on the inflective verb with a potentially regular inflection paradigm, such as: {*ktoś*} *dzwoni zębami* (lit. {*someone*} *rings their teeth*, 'someone feels cold'), {*ktoś*} *pada komuś do nóg* (lit. {*someone*} *falls down to someone else's feet*, 'someone shows their respect towards someone else'), as well as {*komuś*$_{\text{Dat}}$} *dzwoni w uszach* (lit. *it rings in* {*someone's*$_{\text{Dat}}$} *ears*, 'someone has tinnitus'), {*komuś*$_{\text{Dat}}$} *pada na mózg* (lit. *it falls on* {*someone's*$_{\text{Dat}}$} *brain*, 'someone acts irrationally'[1]).

The PRED class consists of units whose verbal component belongs to (primarily) defective verbs, which do not inflect by person and number, only by mode and tense

[1] This type of property is a characteristic feature of inflectional languages, such as Polish, as can be seen in the provided translations. Syntactic complexity, as a result of which the logical subject of the sentence is not expressed in the nominative case but in the dependent case, can only be rendered using literal translation. Equivalent units in English retain the typical canonical syntactic structure in which the subject, performer, or person affected by the state is expressed grammatically in the nominative form.

categories, for example *można*, *należy*, *trzeba* ('can,' 'should,' 'need to'). In grammar studies they are called "modal verbs". As in the case of lexemes, the share of predicative multi-word units in the total number of phrases listed in the dictionary is little. Among over 5,000, only 12 are PRED entries, e.g. {*komuś*/*czemuś*} *można wszystkie żebra policzyć* (lit. *one can count* {*someone's*/*something's*} *ribs*, 'someone/something is thin'), {*komuś*} *brak słów* (lit. {*someone*$_{\text{Dat}}$} *lacks words*, 'someone does not know what to say').

All detailed information about units' paradigms and their limitations are marked at the formal level in graphs.

## *4. Verbel. The Inflection Dictionary of Verbal Phraseological Units*

The theoretical model mentioned in section 2 shows particular steps of language description: from the level of text realisation and interpretation (morphological words), through grouping forms according to their morphosyntactic features (flexemes), to the mental abstractive level in the form of units of language (vocabulas). Therefore, it was decided to implement it in electronic inflectional dictionary of verbal multi-word units. *Verbel* is a digitally born dictionary. Its purpose is to give a full paradigmatic description of multi-word units. It is not the only dictionary of this kind. *Verbel* originates from works related to the description of Polish multi-word units for the purpose of an in-depth analysis of Polish texts and is a continuation of the *SEJF* dictionary (*The Grammatical Lexicon of Polish Multi-Word Expressions*), which is a lexical resource of Polish nominal, adjectival and adverbial multi-word expressions, consisting of about 4,700 multi-word units (Czerepowicka, 2014; Czerepowicka & Savary, 2018). However, the level of complexity of the Polish conjugation system calls for special treatment of verbal words. It turns out to be incompatible with the model used in the *SEJF*, which is simpler and the entry's structure is flat. Consequently, a lexicographic description required a significant reconstruction, which determined the final hierarchical structure of the entry in the *Verbel* dictionary. Since the lexicographic information reflects levels of linguistic description, the dictionary can be applicable in NLP of Polish, such as deep mechanisms of language processing or multi-word units' identification in text. Although it has been compiled with machine processing in mind, it can be useful also for human users.

The dictionary contains over 5,000 verbal multi-word different units, both syntactically and morphologically. The distribution of dictionary entries, including their types mentioned in Section 3, is shown in Table 1. The complexity of the unit's paradigm depends, *inter alia*, on which group the unit belongs to.

| | Units |
|---|---|
| 1st type | 4770 |
| 2nd type | 289 |
| 3rd type | 55 |
| Total | 5,114 |

Table 1. Distribution of verbal types in the *Verbel* dictionary

### 4.1 The structure of a unit

The basic unit in the dictionary is a vocabula, i.e. a unit from the highest level of morphological description. Phraseologisms are listed in the 3rd person singular in the non-past tense if it exists, such as: {*ktoś*/*coś*} <u>*dolewa*</u> *oliwy do ognia*, (lit. {someone/something} is *adding oil to the fire*, 'someone/something is adding fuel to the fire'; {*coś*} <u>*bierze*</u> *w łeb* (lit. *{something} is taking to the head*, 'something, like a plan, is unsuccessful') – imperf; {*ktoś*/*coś*} <u>*doleje*</u> *oliwy do ognia* lit. *{someone/something} will add oil to the fire*, 'someone/something will add fuel to the fire'); {*coś*} <u>*weźmie*</u> *w łeb* (lit. *{something} will take to the head*, 'something, like a plan, will become unsuccessful') – perf.

In line with the Polish lexicographic tradition verbal units should be recorded in the infinitive form. However, there are important reasons to deviate from the known path. The 3rd person form shows the unit in its natural syntactic and semantic context. It also helps to identify a conjugation group, which can be especially useful for human users of the dictionary. This method of lemmatisation was postulated in specialised descriptions (cf. Tokarski, 1973) and has been used in a few Polish dictionaries (cf. Bogusławski & Garnysz-Kozłowska, 1979; Bogusławski & Wawrzyńczyk, 1993; Bogusławski & Danielewiczowa, 2005; Dunaj, 1996).

Beside the type of the unit, each entry gives general and detailed information on the unit. The entry is characterised by an appropriate structure comprised of stages of each units' description, see Fig. 3.

General and detailed information is grouped into particular tabs in the application: general information about the unit on the vocabula level (Opis ogólny hasła, lit. 'general description of an entry'), detailed inflectional information understood as pointing to the main flexeme and a list of all of them (Opis jednostki, lit. 'description of a unit'), a formal description of sub-paradigms in the form of graphs (Opis odmiany fleksemu, lit. 'description of the flexeme's inflection'), forms of particular flexemes (Formy fleksemu, lit. 'forms of flexeme') and paradigms of individual units (Wszystkie formy jednostki, lit. 'all forms of the unit').

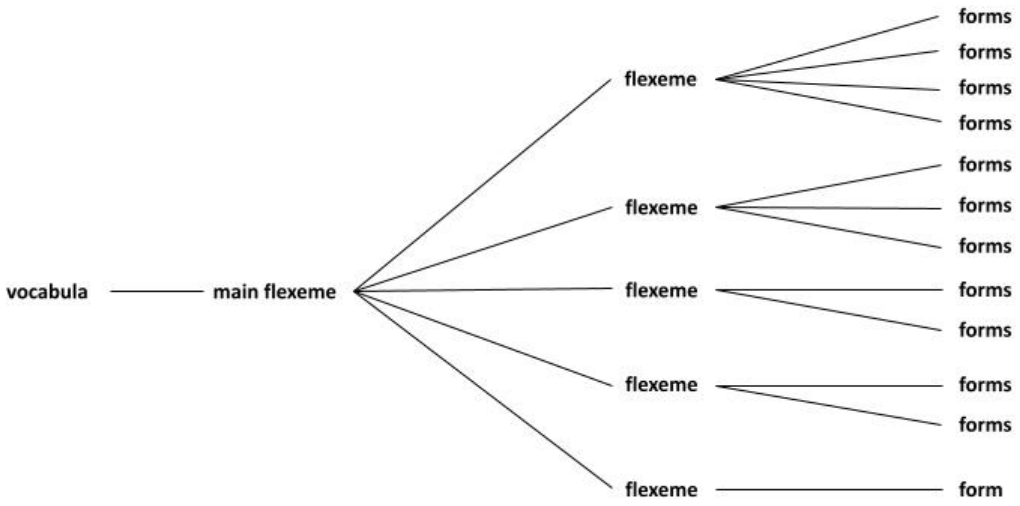



Figure 3. A scheme for the structure of an entry in the *Verbel* dictionary

### 4.2 A general description of the unit

At the initial stage of description, each multi-word verbal unit is assigned to one of two morphological types of vocabulas: verbs (VERB) or predicates (PRED).

In addition to assigning units to a grammatical class, the value of the aspect of phraseologisms is noted – perfective (perf) or imperfective (imperf). The unit's aspect equivalents, if any were determined, are also included here. What is more, this element of an entry presents general descriptive information about the paradigm (F), e.g. full, in the case of defective paradigm, and the types of excluded inflections are provided. It includes other general data about the unit, e.g. possible non-verb variants (W), pragmatic information (P), normative information (N), supplementary grammatical information (G), examples (Np.) and selectively the meaning of the described units:

| | |
|---|---|
| {*ktoś*} *nabiera rumieńców* | *somebody blushes* |
| F: pełny paradygmat | F: full paradigm |
| W: cery, kolorów | W: *lit.* complexion, colours |
| P: tryb rozkazujący w funkcjach wtórnych | P: imperative in secondary functions (a wish, a threat) |

Examples included in entries come from original texts – from NKJP and the resources of the Polish Internet. The shape of a typical entry is shown in Figure 4:

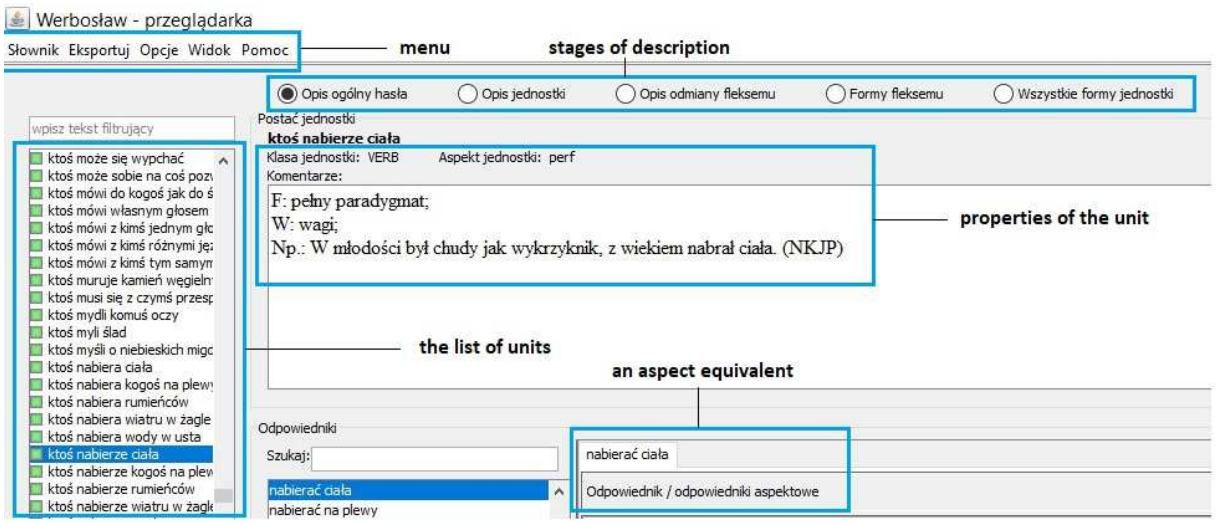



Figure 4. The general description of an entry in the *Verbel* dictionary

Descriptive information about the unit gives an idea of its properties and meaning, additionally illustrating its use in a sentence. This part of the application roughly coincides with the traditional lexicographic description and is advantageous for the human user.

### 4.3 Inflectional information

The following tabs contain more formal inflectional and paradigmatic description of the unit. The next step is to indicate the form of the main flexeme and grammatical characteristics of each component of the multi-word unit. The main flexeme is provided in the infinitive form of the verbal component along with all the lexical parts of the unit, excluding open positions marked with the pronouns *someone*, *something*, e.g. *mieć ręce pełne roboty* (lit. *to have hands full of work* 'to be busy'), *dolać oliwy do ognia* (lit. *to add oil to the fire*, 'to add fuel to the fire'), *pomóc jak umarłemu kadzidło* (lit. *to help like the incense helps the dead*, 'to be of no help at all'). The choice of the infinitive for the base form (main flexeme) was determined by the way forms are created in the dictionary application. They are obtained on the basis of the infinitive form in the morphological generator Morfeusz used in the dictionary (see Woliński, 2014).

Grammatical description of individual components consists of lemmatization and indicating an appropriate morphosyntactic tag. The dictionary provides rudimentary information on the internal syntax of the unit, e.g. by pointing out its main segment – head (Głowa), see Figure 5.

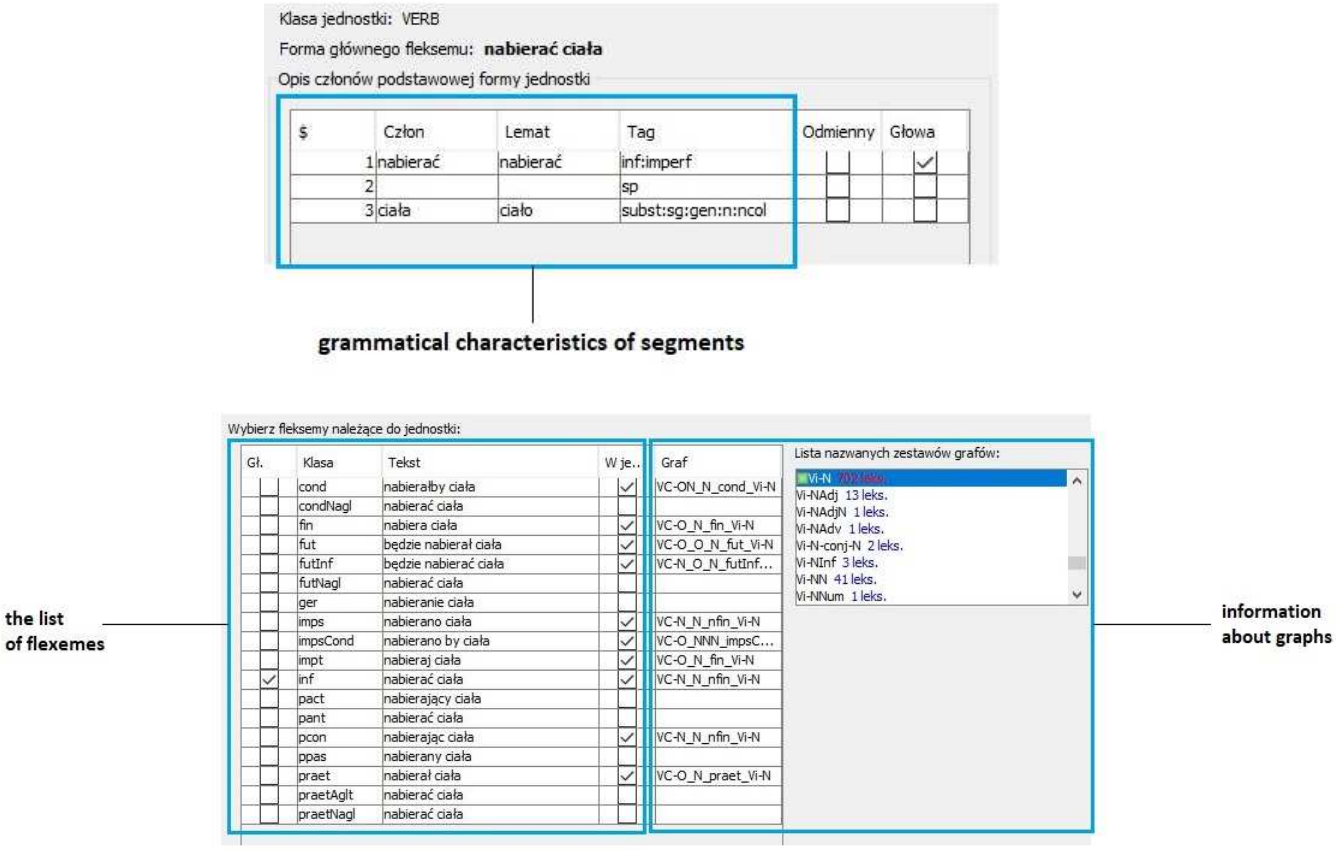



Figure 5. A part of the description of a unit in the *Verbel* dictionary

Then, specific types of flexemes are assigned to each entry. Their number is determined by the value of aspect and specific inflectional features of the unit. For instance, full paradigm imperfective units contain 10 types of paradigmatic words, perfective ones – 8. This tab is crucial for the structure of the entry in the dictionary, as it contains a list of all the flexemes belonging to a given unit, see Figure 5.

### 4.4 Formal description

Generation of the forms of individual flexemes in the dictionary is based on graphs (cf. Marciniak et al. 2011). The relation of a graph to a flexeme is one-sided: each flexeme, regardless of the complexity of its forms, is attached to exactly one graph, but one graph can be assigned to many flexemes which consist of the same number of segments and have the same set of forms. The invariance of units (especially visible in the forms of the past tense, future compound tense, and the conditional mood) is recorded in the form of successive paths in the graph (see Fig. 6). For the purpose of describing over 5,000 phraseological units, 818 individual graphs were created. They contain information about inflectional categories and aspects. The markers of grammatical categories and their values follow the tagset of NKJP.

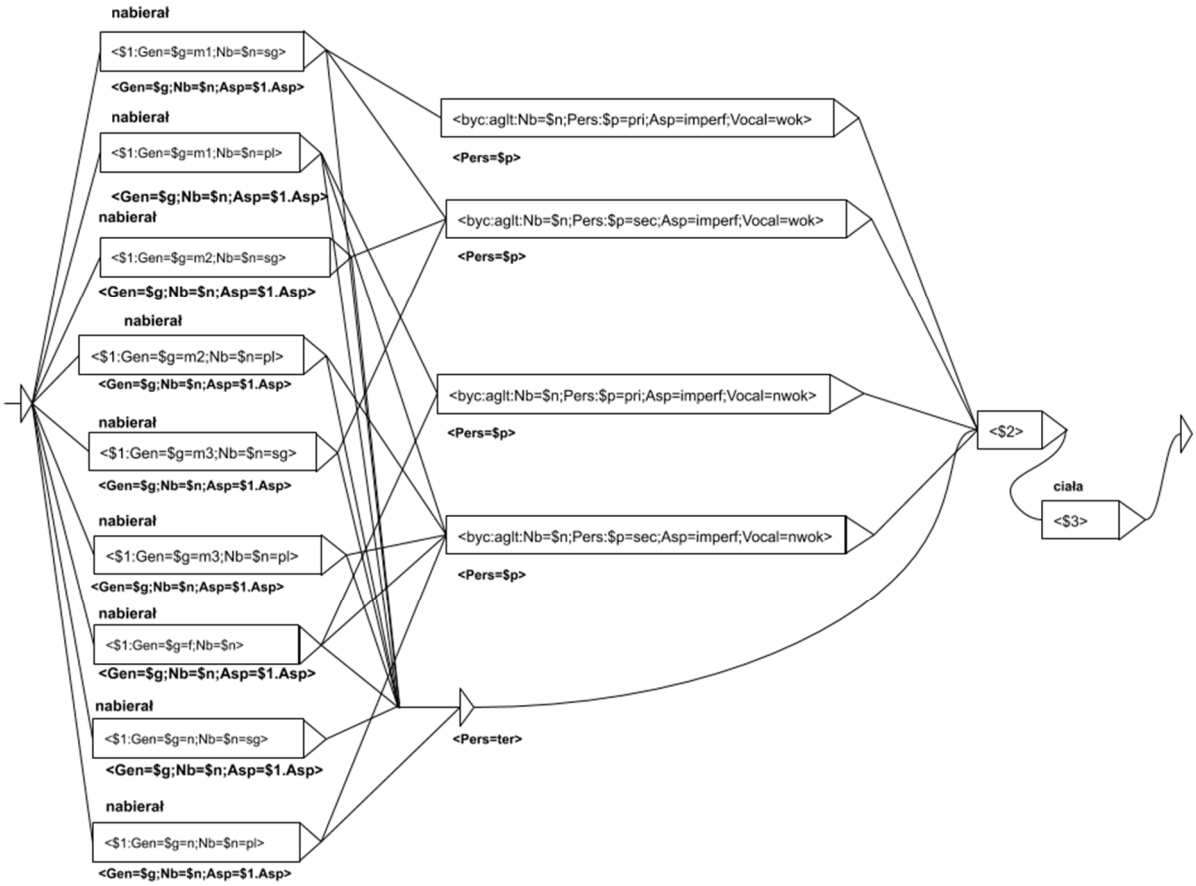



Figure 6. An example of the graph of a past tense flexeme

The graph above presents one of the most complex flexemes – of the past tense. The multitude of paths in the graph is dictated by the complex morphological structure of this type of form. When generating them, several morphosyntactic parameters should be taken into account at the same time, such as person, gender, and vocalism.

Graphs can be grouped into sets on the basis of their morphological and syntactic features. The same set of graphs is assigned to phraseologisms with a similar formal structure and with exactly the same inflectional paradigm. Each set contains a list of flexemes belonging to the unit along with graphs assigned to individual flexemes (Fig.7).

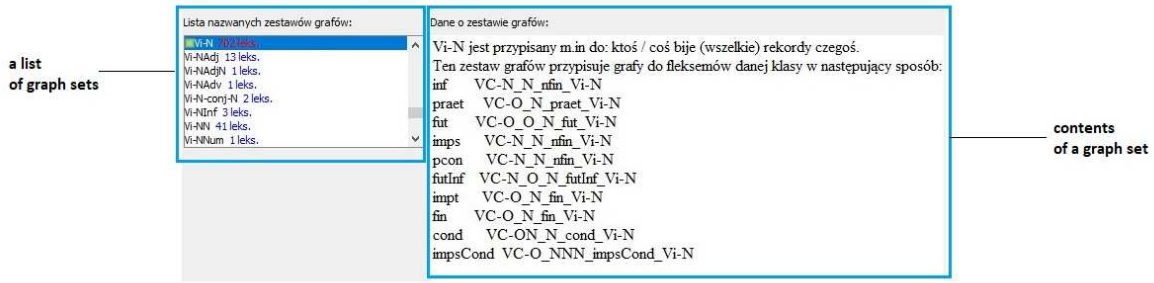



Figure 7. A list of graphs belonging to a Vp-N set

There is a total of 504 graph sets in *Verbel*. Grouping sets allows one to draw conclusions regarding the number of particular syntactic-morphological types of Polish multi-word units. Almost 30% of graph sets concern regular paradigms with a complete set of forms, a vast majority of which belong to imperfective units. Sets that support the greatest number of units have prepositional-nominal or one nominal complement. Respectively, they are attributed to 827 and 768 from over 5000 units. However, a significant part of the sets is needed to create the forms of incomplete, defective paradigms. There are more than 100 sets belonging to single, individual units, such as: {*ktoś*} *przewraca się w grobie* ({*somebody*} *turns* (*over*) *in* (*one's*) *grave*), {*ktoś*} *zjadłby konia z kopytami* ({*somebody*} *could eat a horse including its hooves*), {*ktoś*} *nie dałby za* {*coś*} *złamanego grosza* ({somebody} *doesn't give / won't give single penny for* {*something*}).

### 4.5 A full paradigm

On the basis of graphs, the application generates forms of individual flexemes which constitute separate sub-paradigms. In turn, a set of all sub-paradigms constitutes a complete paradigm of the unit. It is a list of all forms of the unit together with a morphosyntactic tag (Fig. 8).

## 5. Conclusions

Obtaining a full paradigm of a given multi-word unit in the *Verbel* dictionary takes place gradually according to the principle from general to particular, i.e. from general descriptive information about the unit to a list of all its forms (morphological words) with the inflectional characteristics assigned to them. It seems that the data provided at the initial stage (general information about the variant, data about the value of aspect, presence of an aspect equivalent, the meaning and examples) and the final stage (all inflectional forms of the unit) constitute a sufficient lexicographic description.

Placing individual types of information in separate dictionary tabs gives the user the freedom to apply it. It is very likely that an average user will be satisfied with the general description, and perhaps they will also look at the list of forms. On the other hand, a specialist – a linguist, lexicographer, computer scientist – will be curious about various stages of the description and technical ways of their presentation. Although the dictionary contains detailed information described according to a specialised linguistic model, its basic use does not require extensive specialist knowledge. Tabs containing details of the formal description can be omitted without losing the functionality of the dictionary.

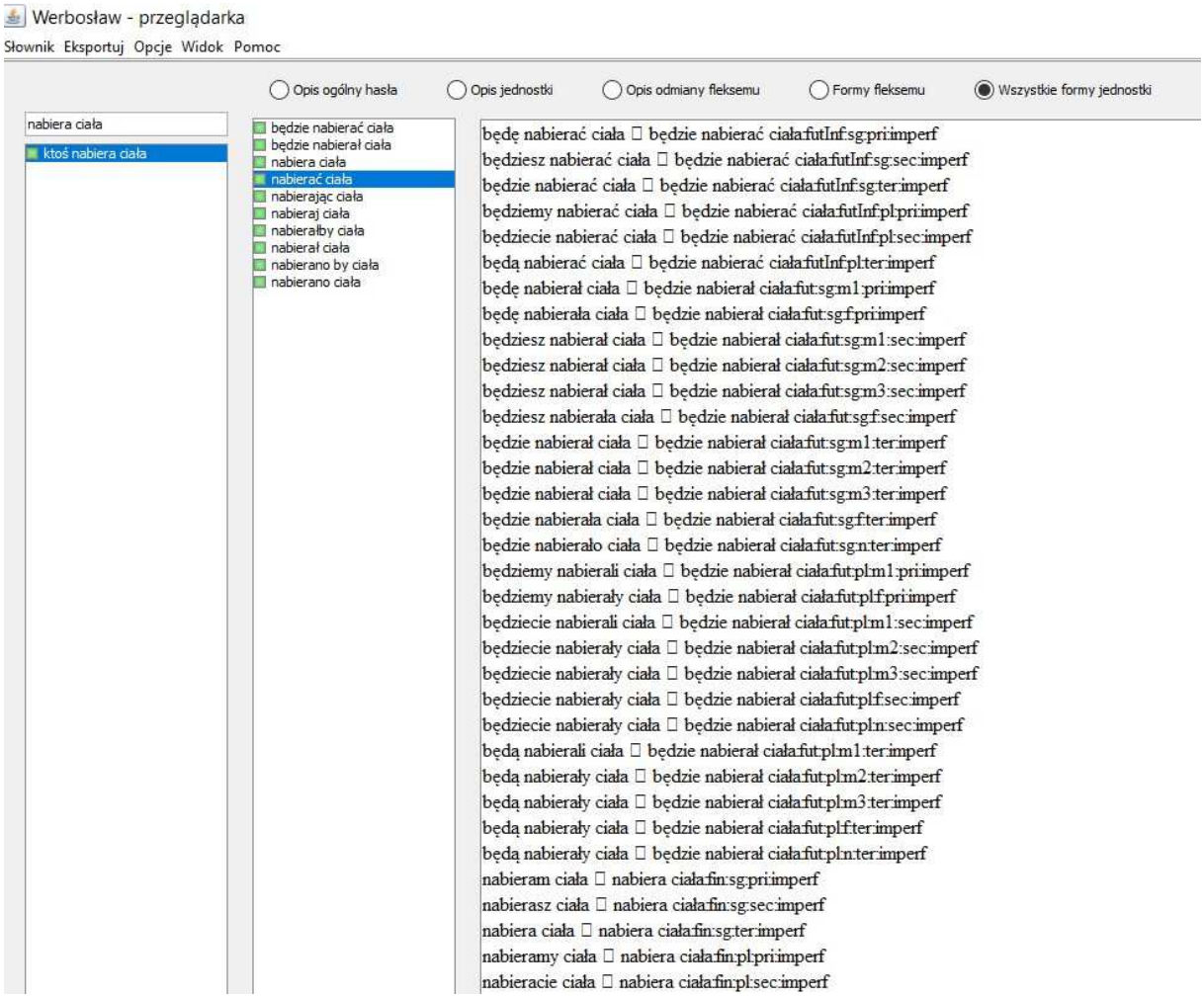



Figure 8. A full paradigm of the unit {*ktoś*} *nabiera ciała* (lit. {*someone*} *gets a body*, 'someone puts on weight')

In this regard, the dictionary may be useful for an average user, not a specialist, especially since the transition from a general description to a full paradigm does not require going through all the description steps in a sequence. Intermediate stages may prove interesting for researchers who focus on describing natural language in a formal manner. In this sense, the dictionary can reach a wide audience. Perhaps it is far from a particularly user-friendly dictionary, and to become one it needs an appropriate interface. Currently, it is an offline resource, and taking into account the expectations of users and technological development the web version would be of greater value. However, these are purely technical conditions. When it comes to the lexicographic layer, the dictionary contains data that is appropriately organised to be a resource for a wide audience.

Since each single form carries a label that includes the name of the base flexeme and its grammatical characteristics (see Figure 8), the data contained in the *Verbel* dictionary can be useful in marking multi-word units in other linguistic tools: text corpora and treebanks, especially because the morphosyntactic marker system used in the dictionary is compatible with the tagset of Polish National Corpus. That is why the dictionary can also be applied in further research on multi-word units in texts.

## 6. References


Baldwin, T. & Kim, S. N. (2010). Multiword Expressions. In: N. Indurhya, F.J. Damerau (eds.) *Handbook of Natural Language Processing*. Boca Raton, USA: CRC Press, pp. 267–292.

Bień, J. S. & Saloni, Z. (1982). Pojęcie wyrazu morfologicznego i jego zastosowanie do opisu fleksji polskiej (wersja wstępna). *Prace Filologiczne*, XXXI, pp. 31–45.

Bogusławski, A. (1976). O zasadach rejestracji jednostek języka. *Poradnik Językowy*, 8(342), pp.356–364.

Czerepowicka, M. (2014). SEJF – Słownik elektroniczny jednostek frazeologicznych, *Język Polski*, XCIV, vol. 2, pp. 116-129.

Czerepowicka, M., Savary, A. (2018). SEJF – a Grammatical Lexicon of Polish Multi-Word Expressions. In: Z. Vetulani, J. Mariani (eds.) *Human Language Technologies as a Challenge for Computer Science and Linguistics: 7th Language and Technology Conference, LTC 2015, Poznań, Poland, November 27–29 2015, Revised Selected Papers, In memoriam Adam Kilgarriff*. Lecture Notes in Artificial Intelligence, vol. 10930. Berlin: Springer-Verlag, pp. 59–73.

Iosad, P., Koptjevskaja-Tamm, M., Piperski, A. & Sitchinava, D. (2018). Depth, brilliance, clarity: Andrey Anatolyevich Zaliznyak (1935–2017). *Linguistic Typology*, 22(1), pp. 175–184.

Kosek, I., Czerepowicka, M., Przybyszewski, S. (2020). *Verbel. Elektroniczny słownik paradygmatów polskich frazeologizmów czasownikowych. Teoria, problemy, prezentacja*. Olsztyn: University of Warmia and Mazury.

Lewicki, A. M. (1986). Składnia związków frazeologicznych. *Bulletin de la Société Polonaise de Linguistique*, XL, pp. 75–83.

Marciniak, M., Savary, A., Sikora, P. & Woliński, M. (2011). Toposław – a lexicographic framework of multi-word units. In: Z. Vetulani (ed.) *Human Language Technology. Challenges for Computer Science and Linguistics: 4th Language and Technology Conference, LTC 2009, Poznań, Poland, November 6-8, 2009, Revised Selected Papers*. Lecture Notes in Artificial Intelligence, vol. 6562. Berlin: Springer-Verlag, pp. 139–150.

Mel'čuk, I. (2006). Explanatory Combinatorial Dictionary. In: G. Sica (ed.) *Open Problems in Linguistics and Lexicography*. Monza: Polimetrica, pp. 225–355.

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

Sag I.A., Baldwin T., Bond F., Copestake A. & Flickinger D. (2002). Multiword Expressions: A Pain in the Neck for NLP. In: Gelbukh A. (eds.) *Computational Linguistics and Intelligent Text Processing.* CICLing 2002. Lecture Notes in Computer Science, vol. 2276. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-45715-1_1.

Woliński, M. (2014). Morfeusz reloaded. 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, pp. 1106–1111, Reykjavík, Iceland, 2014. ELRA.

Tokarski, J. (1973). *Fleksja polska*. Warsaw: Państwowe Wydawnictwo Naukowe.

**Dictionaries & Websites:**

Bogusławski, A. & Danielewiczowa, M. (2005). *Verba polona abscondita. Sonda słownikowa III*. Warsaw: Elma Books.

Bogusławski, A. & Garnysz-Kozłowska, T. (1979). *Addendum to Polish phraseology. An introductory issue*. Edmonton: Linguistic Research.

Bogusławski, A. & Wawrzyńczyk, J. (1993). *Polszczyzna, jaką znamy. Nowa sonda słownikowa.* Warsaw: University of Warsaw.

Dunaj, S. (ed.). (1996). *Słownik współczesnego języka polskiego.*Warsaw: Wilga.

Mel'čuk, I. & Zholkovsky, A. (1984). *Explanatory Combinatorial Dictionary of Modern Russian. Semantico-syntactic Studies of Russian Vocabulary*. Viena: Wiener Slawistischer Almanach.

NKJP: Narodowy Korpus Języka Polskiego. Accessed at: http://nkjp.pl/ (20 May 2021).

Saloni, Z. (2007). *Czasownik polski. Odmiana. Słownik 12 000 czasowników*. Edition III, changed. Warsaw: Wiedza Powszechna. Accessed at: https://depot.ceon.pl/handle/123456789/20067 (27 May 2021).

SEJF: *Słownik elektroniczny jednostek frazeologicznych*. Accessed at: http://zil.ipipan.waw.pl/SEJF (20 May 2021).

SGJP: *Słownik gramatyczny języka polskiego*. Accessed at: http://sgjp.pl/ (20 May 2021)

Verbel: *Verbel. Elektroniczny słownik paradygmatów polskich frazeologizmów czasownikowych*. Accessed at: http://uwm.edu.pl/verbel/ (20 May 2021).

Zaliznyak, A. A. (1977). *Grammatičeskij slovar' russkogo jazyka. Slovoizmenenie* [*A grammatical dictionary of Russian: Inflection*]. Moscow: Russkij jazyk.