RuSkELL: Online Language Learning Tool for Russian Language

Valentina Apresjan, Vít Baisa, Olga Buivolova,
Olga Kultepina, Anna Maloletnjaja
National Research University “Higher School of Economics”, School of Linguistics
Vinogradov Russian Language Institute
NLP Centre, Masaryk University
e-mail: vapresyan@hse.ru, xbaisa@fi.muni.cz, ovbuyvolova@edu.hse.ru,
oakultepina@edu.hse.ru, maloletnyaya@gmail.com

Abstract

RuSkELL ("Russian + Sketch Engine for Language Learning") is a new online resource intended for researchers and learners of Russian. It incorporates a specially pre-processed corpus and the interface which allows users to search for phrases in sentences, extract salient collocates and show similar words. The tool builds upon its English counterpart SkELL (Baisa & Suchomel 2014). The aim of the project is to adapt the existing SkELL tool to Russian, improve its performance and make it user-friendly to Russian users. The existing problems include errors in query output and insufficiently transparent interface. The project aspires to solve them by 1) modifying Sketch grammar rules to exclude irrelevant output and to add informative collocations unaccounted for in the existing Sketch grammar; 2) providing collocation groups with easy-to-understand labels in Russian. We describe the process of building the language data and problems we need to address to accommodate the tool for the specificities of the Russian language.

Keywords: online language tool; Sketch Engine for Language Learning; sketches; collocations

1. RuSkELL Corpus[1]

The corpus contains texts downloaded by web crawler SpiderLing (Suchomel, Pomikálek, 2012) from the Russian Internet in 2011. The seed URLs (starting points for the crawler) were obtained from web search engines using the Corpus Factory method (Kilgarriff et al. 2010). The following procedures for cleaning web content were applied to 4 TB of HTML data: 1) character encoding detection, 2) language identification, 3) boilerplate removal, 4) near duplicate paragraphs removal. The size of the original corpus data is 20.2 billion tokens in 198 GB of plain text. 99.8% of documents come from the Russian national top level domain.ru There are 405,748 web domains represented by at least one document in the corpus. The most frequent sites are kontrolnaja.ru, news.yandex.ru, alterauto.ru, pressarchive.ru and com.sibpress.ru covering just 0.09% of all documents. The corpus was further cleaned of obscene language, using a list of words, prohibited in .рф domain space[2]. All sentences were sorted according to a special GDEX (Good Dictionary EXamples) score (Kilgarriff et al., 2008), favouring average-long sentences with mid-frequency words which are more suitable for learners and only top 68,232,088 sentences were used, yielding the corpus with 975,742,959 words.

2. The Interface[3]

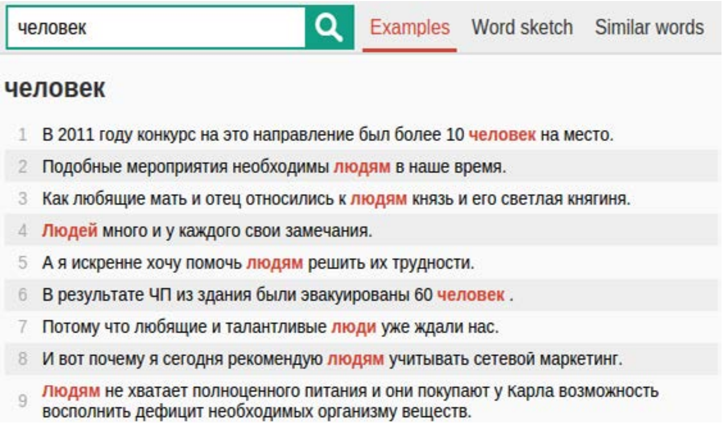
The interface was designed to be intuitive[4]. It offers three functions: examples (full-text search), word sketch (collocation profiles) and similar words (distributional thesaurus). See Figure 1 for an example of sentences for “человек” and Figure 2 for an example of similar words.

[1] ruskell.sketchengine.co.uk

[2] The corpus cleaning was executed by Timur Iskhakov under the mentorship of Andrey Shestakoff and Ekaterina Chernyak (Faculty of Computer Science, National Research University “Higher School of Economics”).

[3] http://corpus.tools

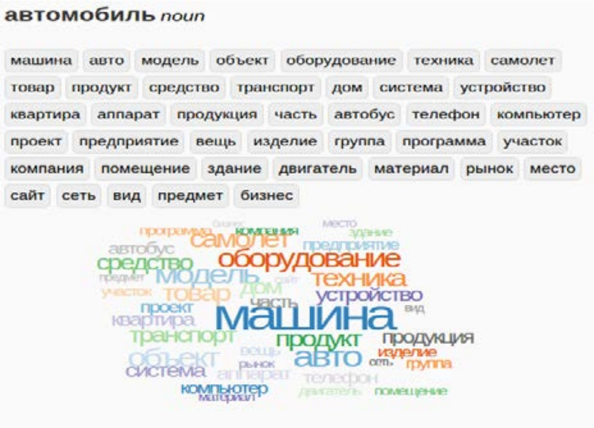
[4] A complete guide to Sketch Engine and SkELL may be found in (Thomas 2016).



человек

1 В 2011 году конкурс на это направление был более 10 человек на место.
2 Подобные мероприятия необходимы людям в наше время.
3 Как любящие мать и отец относились к людям князь и его светлая княгиня.
4 Людей много и у каждого свои замечания.
5 А я искренне хочу помочь людям решить их трудности.
6 В результате ЧП из здания были эвакуированы 60 человек .
7 Потому что любящие и талантливые люди уже ждали нас.
8 И вот почему я сегодня рекомендую людям учитывать сетевой маркетинг.
9 Людям не хватает полноценного питания и они покупают у Карла возможность восполнить дефицит необходимых организму веществ.

Figure 1. Example sentences for query “человек” containing various word forms.

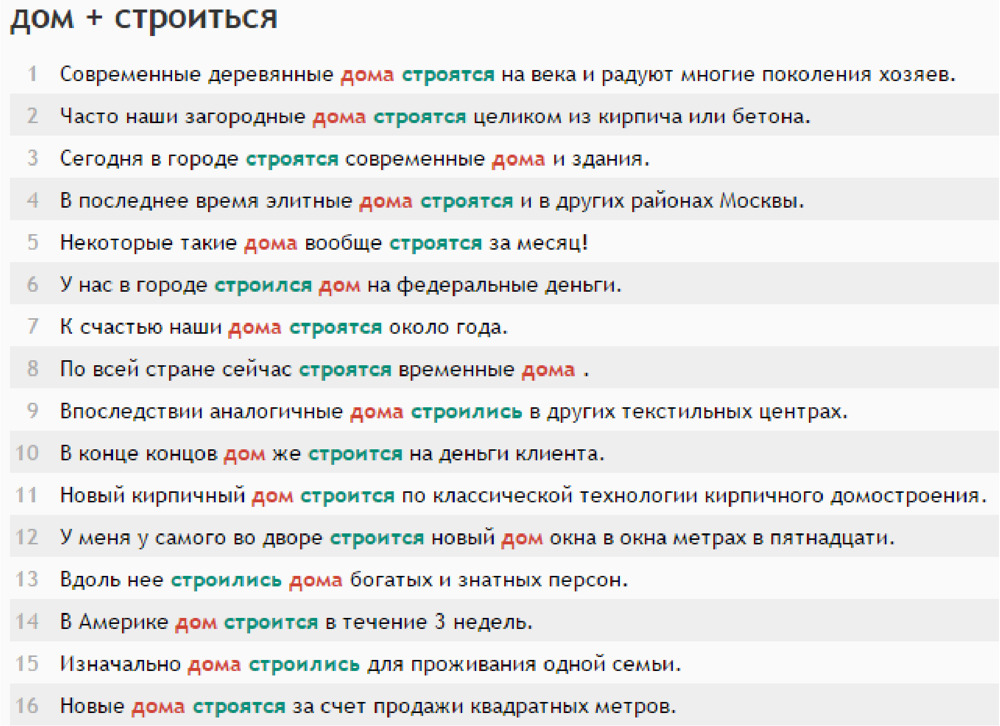


автомобиль *noun*

машина авто модель объект оборудование техника самолет
товар продукт средство транспорт дом система устройство
квартира аппарат продукция часть автобус телефон компьютер
проект предприятие вещь изделие группа программа участок
компания помещение здание двигатель материал рынок место
сайт сеть вид предмет бизнес

Figure 2. Similar words for “автомобиль”.

If you click on a similar word in the thesaurus, you get a word sketch for that word. If you click on a collocate, you see a concordance with the headword and the collocation highlighted (see Figure 3 for an example).



дом + строиться

1 Современные деревянные дома строятся на века и радуют многие поколения хозяев.
2 Часто наши загородные дома строятся целиком из кирпича или бетона.
3 Сегодня в городе строятся современные дома и здания.
4 В последнее время элитные дома строятся и в других районах Москвы.
5 Некоторые такие дома вообще строятся за месяц!
6 У нас в городе строился дом на федеральные деньги.
7 К счастью наши дома строятся около года.
8 По всей стране сейчас строятся временные дома .
9 Впоследствии аналогичные дома строились в других текстильных центрах.
10 В конце концов дом же строится на деньги клиента.
11 Новый кирпичный дом строится по классической технологии кирпичного домостроения.
12 У меня у самого во дворе строится новый дом окна в окна метрах в пятнадцати.
13 Вдоль нее строились дома богатых и знатных персон.
14 В Америке дом строится в течение 3 недель.
15 Изначально дома строились для проживания одной семьи.
16 Новые дома строятся за счет продажи квадратных метров.

Figure 3. Concordance with highlighted collocates for “дом +строиться”.

3. Advantages – Discovering Senses with Word Sketches

The usefulness of RuSkELL in lexicography and language teaching is apparent. A word sketch (Kilgarriff et al., 2004) search provides exhaustive information about a word's polysemy, collocations, idioms, and even colligations, or grammatical properties specific to certain senses of lexical items (Atkins, Rundell 2008). Consider Russian verb *pojti* ('to go, to start going'). First of all, relations post_inf and subject (that give collocations with the most frequent infinitives following *pojti* and the most frequent nouns which act as subjects of *pojti*) allow one to establish the approximate list of senses for this highly polysemous word. Consider some of them:

- 'to go, to start walking somewhere with some purpose', as in *pojti guljat'* 'to go for a walk', *pojti igrat'* 'to go to play';
- to start (about activities)', as in *Igra poshla* 'The game started';
- 'to start (about processes)', as in *Poshel dozhd'* 'Rain started';
- 'to start flowing', as in *Poshla krov'* 'Blood started flowing';
- 'to spend', as in *Den'gi poshli na chto-to* 'Money was spent on smth.';
- 'to originate', as in *Otsjuda poshlo nazvanie* 'Name originates from this'.

Moreover, examples with nouns from the subject list reveal information about syntactic peculiarities of *pojti* in different senses: thus, in senses 3 and 4 the verb is usually in preposition to the subject. This inversion is typical of existential predicates in Russian, cf. *Zdes' vodjatsja oleni* 'Deer harbor here', literally 'Here harbor deer-NOM'. This is a language-specific feature and therefore useful to know for a language learner, yet not something which could easily be found in a learner's grammar or in a regular dictionary. In this respect, RuSkELL allows one to extract information on colligations that is far from trivial.

Further inspection of collocations yields more senses; e.g. object3 reveals a figurative meaning 'to satisfy', as in *pojti navstrechu chjim-to zhelanijam* 'to satisfy one's wishes', literally 'to go towards someone's wishes-DAT'.

Inst_modifier produces yet new senses, again with non-trivial syntactic peculiarities, namely with an instrumental in the semantic role of activity, agent, or result: 'to start military actions against smb.', as in *pojti vojnoj* 'to start a war against smb., lit. 'to go war-INSTR', 'to enlist in a certain capacity', as in *pojti dobrovol'cem* 'to volunteer', lit. 'to go volunteer-INSTR', 'to start getting deformed in a certain way', as in *pojti treshchinami* 'to start cracking', lit. 'to go cracks-INSTR'.

Inst_modifier and prepositional collocations also uncover a wealth of idioms, e.g. *pojti praxom* 'to go down the tubes', literally 'to go ash-INSTR', *pojti v goru* 'to hit it big', literally 'to go into the mountain' and many others.

4. Issues

Using RuSkELL may also be difficult or confusing, and one of the objectives of the current project is to identify the problems and suggest ways of minimizing them. This objective is achieved in the following steps: 1) testing RuSkELL, 2) identifying mistakes, sources of confusion and gaps in the outcome, 3) suggesting alterations and 4) testing altered search on users.

Some of the problems are easily observable through a "surface scratch" of RuSkELL functions. For example, the names of grammatical relations that underlie the collocation groups (object2, object3, object4, instr_modifier, verb_post_inf) are far from transparent.

These are relatively easy to fix because they are not triggered by inherent peculiarities of the Russian language. However, further and deeper problems arise with more testing. At the current stage, the following problems have been identified:

- confusion of object2 (genitive complement) and object4 (accusative complement) relations. For example, for the verb *vesti* 'to lead' which normally governs Patient in the accusative, we get the noun *vojna* 'war' both in object2 and object4 lists; for the verb *ljubit'* 'to love' which normally governs object in the accusative, we get the noun *rebenok* both in object2 and object4. The confusion is partly due to accusative-to-genitive change in verbs under negation, and partly to the homonymy of case forms. Same causes bring about the confusion of object2_of and object4_of relations;

- confusion between object2 (genitive complement) and object3 (dative complement) relations, as well as between their dual relations object2_of and object3_of. For example, for the verb *kasat'sja* 'to touch' which only governs Patient in the genitive, we get nouns *grud'* 'chest', *istorija* 'story', *poverxnost'* 'surface' both in object2 and object3 list. Here again the source of confusion lies in the morphological homonymy (between the forms of genitive and dative for feminine nouns of the third declension type in singular);
- confusion of subject (nominative subject) and object4 (accusative complement) relations, as in *sleduet pereryv* 'follows break-NOM', which instead of subject, is wrongly identified as object4; the cause of which is homonymy of the forms of nominative and genitive inanimate for masculine nouns in singular and plural and feminine nouns of the first declension type in plural;
- confusion of object3 (dative complement) and object4 (accusative complement) relations, as in *sledovat' tradicii* 'to follow tradition-DAT-SING', which instead of object3, is wrongly identified as object4, due to homonymy of the forms of dative singular and accusative plural for feminine nouns of the third declension type.

At the present stage of our project, it has proved impossible to improve the situation with grammatical homonymy, so it remains subject to further experiment. One of the possible avenues in the reduction of homonymy is the modification of Sketch grammar rules in order to exclude negated verbs from consideration, as they considerably affect the statistics of genitive-accusative confusion. Discrimination between genitive and accusative complements might also be aided by excluding Russian verbs that require genitive complements and cannot take accusative complements from the output in the Obj4 group (provided it is technically possible). Barring negation, genitive verbs form a rather compact morphosyntactic class in Russian; displayed below is a manually compiled table of the most frequent verbs taking genitive complements.

Table 1. The most frequent Russian verbs with genitive complements.

Genitive verbs without preposition	**Freq**	**Genitive verbs with a preposition**	**Freq**
Stoit' 'be worth'	501.9	*Sprosit' u* 'ask'	573.9
Kasat'sja 'touch'	154.7	*Ujti iz/ot* 'go away'	315.6
Dostignut' 'achive'	97.4	*Sostojat' iz* 'consist of'	139.9
Derzhat'sja 'hold'	86.4	*Ujexat' iz/ot* 'move away'	117.4
Dobit'sja 'reach'	69.5	*Otkazat'sja ot* 'abandon'	115.8
Dobivat'sja 'obtain'	29.4	*Zaviset' ot* 'depend on'	115
		Otlichat'sja ot 'differ from'	98.9
		Dostat' do 'reach, touch'	94.4
		Dojti do 'come to'	88.6
		Isxodit' ot 'come from, issue from'	69
		Otojti ot 'pull out'	55.9
		Sojti s 'come down from'	55.4
		Dobrat'sja do 'reach, get to'	49.6
		Obrazovat'sja iz 'arise from'	42.1
		Vyskochit' iz 'jump out'	42.1
		Skladyvat'sja iz 'turn out'	41.4
		Vyrvat'sja iz 'break free'	31.8
		Obojti vokrug 'pass around'	30.4

		Vybrat'sja iz 'get out	29.2
		Skryt'sja ot 'hide oneself from'	28.7
		Izbavit'sja ot 'get rid of'	27
		Uderzhat'sja ot 'refrain'	26

There are, however, errors in the output which are due to easily fixed deficiencies in the Sketch grammar; these have been repaired in the current version of RuSkELL. They are as follows:

- noise in object2_of relations, as in the word sketch of *chelovek* 'human, person' where the list of verbs includes intransitive verbs such as *rabotat'* 'to work', *spat'* 'to sleep', *zhit'* 'to live' which cannot possibly govern *chelovek* as they do not take complements. The analysis of examples shows that in all such examples the verb is in the participle form and therefore does not govern the noun, but has, in contrast, attributive function with respect to it: *zhivushchix ljudej* 'living people', *rabotajushchix ljudej* 'working people'. This problem is fixed by excluding collocations with participles from the rule;
- confusion of and/or relation and post_inf relation for verbs, as in the sketch of the verb *pojti* 'to go, to start going' where both relations produce the verb lists including *guljat'* 'to go for a walk', *spat'* 'to sleep', *rabotat'* 'to work'. In fact, all verbs listed in both these categories only belong to the second one. They represent not the syntactic relation of coordination implied by and/or, but the relation of government implied by post_inf. The verb *pojti* 'to go' in Russian belongs to the group of verbs that take infinitival complements like modals: *pojti spat'* 'to go sleep', *pojti rabotat'* 'to go work'. The problem is solved by changing the rule as to include a conjunction ('and' or 'or') between the verbs.

During the test process we have also identified certain gaps in the output that require the addition of new grammatical relations and rules to the Russian Sketch grammar. For example, while the present version of RuSkELL provides rich sketches for verbs, nouns and adjectives, other parts of speech are scarcely represented. We suggest certain additions to the Sketch grammar that improve sketches for adverbs and add sketches for numerals. We also include additions to verb sketches that are aimed at reflecting collocational properties of certain constructions which are peculiar to Russian, such as, for example, depictives (e.g. 'to remember somebody young-INSTR'). The aim of these alterations is to advance the search in RuSkELL by tailoring the rules to the peculiarities of the Russian grammar. The changes are reflected in the experimental version of RuSkELL (http://ruskell.sketchengine.co.uk/run2.cgi/skell). The additions to the Russian Sketch grammar and the rationale behind them are discussed below.

5. Filling in the Gaps in Grammatical Relations

As stated above, certain adjustments have been made in the Russian Sketch grammar to improve the outcome.

One of the problems we faced concerned the relation adv_modifies. In the present development version of RuSkELL the group adv_modifies contains both verbs and adjectives modified by an adverb. Clearly, it is not a particularly user-unfriendly solution. In our experimental version, we divided this group into two: adv_modifier_verb (*ploxo spat'* 'to sleep badly') and adv_modifier_adj (*tochno uverenniy* 'absolutely sure'). Two separate groups are not only more user-friendly, they are also a more natural solution from the syntactic point of view. Moreover, the adv_modifier_adj group also returns useful collocations with participles modified by adverbs; e.g. in the upgraded version of RuSkELL, the adverb *ploxo* ('badly') has a collocation with participle *organizovanniy* ('organized').

Futhermore, this solution increases the output for each grammatical relation: there are 15 collocations for each part of speech (in total 30) instead of 15 mixed verb-adjective collocations as before.

Another proposed addition to the Sketch Grammar concerns depictive construction, i.e. verbs followed by adjectives in the instrumental case (‘*I saw him drunk*-INSTR’, ‘*He looks tired*-INSTR, *She seems old*-INSTR’). This construction is possible only for certain semantic classes of verbs and adjectives. The knowledge of its typical colexification is an essential part of Russian language competence and therefore useful for language learners.

6. Addition of New Parts of Speech: Numerals

When adding new parts of speech to the Sketch grammar, we were guided by various considerations. One of them was the frequency of collocations, and from this point of view, numerals were indicated as an informative addition. In the experimental version of RuSkELL, we included rules that find collocations with ordinal and cardinal numerals.

Ordinal numerals have only one collocation group – num_modifies, which returns collocations such as *vtoroj etazh* ‘second floor’, *vtoraja polovina* ‘second half’, *vtoroj tajm* ‘second round’.

Cardinal numerals are presented by two groups: num_object2_of and num_inst. The first group contains collocations such as *dva goda* ‘two years’ *dva raza* ‘two times’, *dva chasa* ‘two hours’, *dva desjatka* ‘two dozens; lit. two tens’; the second group contains collocations such as ‘dvumja rukami’ ‘with both hands’, *dvumja rjadami* ‘in two rows’.

7. Making RuSkELL More User-friendly

In order to make RuSkELL easier to use, collocation groups have been assigned more transparent headings. Instead of the technical terms for grammatical relations (such as object2, object4 etc.), each grammatical relation was given a heading reflecting its syntactic nature in the terms that are either widely used in learner’s grammars and therefore familiar to a regular user or that are self-explanatory. At the first stage of the experiment, the headings were tested on a sample group of users (27 people including native speakers and language learners). Our main criterion in renaming was user-friendliness for the target audience. Since RuSkELL is orientated towards Russian language learners of different levels, we want include apart from the Russian translations, their English equivalents as well. For example object2 has been renamed as “Дополнение в родительном падеже” / “Genitive complement”. The whole list of new names is in Table 2 in the Appendix.

8. Conclusion

This paper introduces a novel language resource for Russian language learners. It is based on a very large Russian corpus which was processed by the state-of-the-art tools and sorted by GDEX score to favour simple sentences suitable for studying language phenomena. We identified several problems and suggested a way of fixing them to limit a number of possible errors in the data presented to the users and to provide useful collocations in the output. These measurements were applied and the tool is now publicly available in its test version.

The interface will be free for anyone who wants to study Russian language via examples from real language. We believe that this service will be a useful accompanying tool for language teachers and their students.

References

Atkins, S.B.T., Rundell, M. (2008). *The Oxford guide to practical lexicography*. Oxford University Press.

Baisa, V. & Suchomel, V. (2014). SkELL: Web Interface for English Language Learning. In *Eighth Workshop on Recent Advances in Slavonic Natural Language Processing*.

Kilgarriff, A., et al. (2004). Itri-04-08 the sketch engine. In *Information Technology105 (2004): 116.*

Kilgarriff, A., et al. (2008). GDEX: Automatically finding good dictionary examples in a corpus. *Proceedings of the XIII EURALEX International Congress (Barcelona, 15-19 July 2008)*.

Kilgarriff, A., Reddy S., Pomikálek, J. & Avinesh, P.V.S. (2010). A Corpus Factory for Many Languages. In *LREC*.

Kilgarriff, A., et al. (2015). *Longest–commonest Match*.

Suchomel, V., Pomikálek, J. (2012). Efficient web crawling for large text corpora. In *Proceedings of the seventh Web as Corpus Workshop (WAC7)*, pp. 39-43.

Thomas, J. (2016). The 2nd edition of Discovering English with Sketch Engine (DESKE).

Acknowledgements

The funding was provided by the following grant: the Programme for fundamental research of the Russian Academy of Sciences, “Russian Lexicon as an element of Russian culture: systematic organization of the Lexicon and its reflection in the dictionary”, 2015-2017. This work has also been partly supported by the Ministry of Education of CR within the LINDAT-Clarin project LM2015071 and the Norwegian Financial Mechanism 2009-- 2014 and the Ministry of Education, Youth and Sports under Project Contract no.~MSMT- 28477/2014 within the HaBiT Project 7F14047.

Appendix

Table 2: Renaming of grammatical relations.

Previous names	New Russian names	New English names
subject of %w/verbs with %w as subject	подлежащее при %w/глагол с %w в роли подлежащего	subject of %w / verbs with %w as subject
object2/object2_of	дополнение в родительном падеже при %w / глаголы с %w в роли дополнения в родительном падеже	genitive complement of %w / verbs with %w as genitive complement
object3/object3_of	дополнение в дательном падеже при %w / глаголы с %w в роли дополнения в дательном падеже	dative complement of %w / verbs with %w as dative complement
object4/object4_of	дополнение в винительном падеже при %w / глаголы с %w в роли дополнения в винительном падеже	accusative complement of %w/ verbs with %w as accusative complement
inst_modifier/inst_modifies	дополнение в творительном падеже при %w / глаголы с %w в роли дополнения в творительном падеже	instrumental complement of %w / verbs with %w as instrumental complement
gen_modifier/gen_modifies	%w подчиняет существительное в родительном падеже / %w подчиняется существительному в родительном падеже	genitive modifier of % / nouns with %w as genitive modifier
a_modifier/modifies	определение при %w/ существительное	adjective modifier of %w/ nouns

	с %w в роли определения	with %w as adjective modifier
adv_modifier / adv_modifies	обстоятельство при %w / глаголы с %w в роли обстоятельства	adverbial modifier of %w/ verbs with %w as adverbial modifier
adv_modifier_verb / adv_modifies_verb	обстоятельство при %w / глаголы с %w в роли обстоятельства	adverbial modifier of %w/ verbs with %w as adverbial modifier
adv_modifier_adj/adj_modifies_adv	обстоятельство при %w / прилагательное с % в роли обстоятельства	adverbial modifier of %w/ /adjectives with %w as adverbial modifier
num_object2_of	%w управляет существительным в родительном падеже	nouns in genitive with %
num_inst	%w согласуется с существительным в творительном падеже	nouns with %w with modifier in instrumental case
num_modifies	существительное с %w в роли определения	nouns with %w as modifier
быть_adj/subj_быть	прилагательное в функции сказуемого при %w / существительные в роли подлежащего с %w в роли сказуемого	adjective as predicate with %w / nouns as subject with %w as predicate
modal_inf/modal	инфинитивы при %w / модальное слово при %w	infinitives with %w / modal with %w
post_inf/verb_post_inf	%w перед инфинитивом / %w после инфинитива	infinitives after %w / verbs followed by %w in infinitive
prec_prep	предлоги перед %w	prepositions before %w
post_prep	предлоги после %w	%w after preposition
passive/subj_passive	глаголы в пассиве при %w / существительные с %w в пассиве	verbs in passive with %w / nouns with %w in passive
pp_%(3.lemma)	предлог %(3.lemma) после %w	preposition %(3.lemma) after %w
pp_obj_%(3.lemma)	существительное в составе предложной группы с предлогом %(3.lemma) при %w	nouns in prepositional phrases with %(3.lemma) with %w