Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 2788–2797
Marseille, 20-25 June 2022

Constructing a Lexical Resource of Russian Derivational Morphology

Lukáš Kyjánek[◇], Olga Lyashevskaya[•], Anna Nedoluzhko[◇], Daniil Vodolazsky[⊖], Zdeněk Žabokrtský[◇]
[◇]Charles University, Faculty of Mathematics and Physics, Prague, Czech Republic
[•]National Research University Higher School of Economics, Moscow, Russia
[⊖]Sber, Moscow, Russia
{kyjanek, nedoluzko, zabokrtsky}@ufal.mff.cuni.cz olesar@yandex.ru daniil.vodolazsky@mail.ru

Abstract

Words of any language are to some extent related thought the ways they are formed. For instance, the verb *exempl-ify* and the noun *example-s* are both based on the word *example*, but the verb is derived from it, while the noun is inflected. In Natural Language Processing of Russian, the inflection is satisfactorily processed; however, there are only a few machine-trackable resources that capture derivations even though Russian has both of these morphological processes very rich. Therefore, we devote this paper to improving one of the methods of constructing such resources and to the application of the method to a Russian lexicon, which results in the creation of the largest lexical resource of Russian derivational relations. The resulting database dubbed DeriNet.RU includes more than 300 thousand lexemes connected with more than 164 thousand binary derivational relations. To create such data, we combined the existing machine-learning methods that we improved to manage this goal. The whole approach is evaluated on our newly created data set of manual, parallel annotation. The resulting DeriNet.RU is freely available under an open license agreement.

Keywords: Russian, language resource, derivational network, derivational morphology, machine learning

1. Introduction

Russian is a language with rich derivational morphology (Štekauer et al., 2012; Körtvélyessy, 2016) and yet there are only a few resources and tools to process it compared to those for inflectional morphology.[1] Indeed, in Natural Language Processing, the two morphological processes are usually kept apart. They are distinguished on the basis of their regularity and meaning of affixes that are attached to so-called morphological bases when coining inflected or derived words. First, the inflectional morphology seems to be more regular than derivational morphology. Second, an attached affix in the process of inflection/derivation conveys different meaning (grammatical/lexical), see Lipka (2010, p. 71). From this perspective, inflection is the main source of word forms because it conveys grammatical categories, such as cases and tenses, by attaching grammatical affixes, while derivation is the main source of new lexemes as it changes the lexical meaning by attaching lexical affixes. For instance, the noun учителя '*uchitelya*' (*teachers*) is inflected by attaching the inflectional affix -я '*-ya*' to the morphological base of the noun учитель '*uchitel*' (*teacher*), which is derived by attaching the lexical affix -тель '*-tel*' to the morphological base of the verb учить '*uchit*' (*to teach*).[2]

As a consequence of the lower paradigmaticity of derivations, the creation of resources for derivational morphology is complicated. The common methods to build these resources exploit regular expressions or pattern-matching techniques using which they search for derivational relations between words from a given lexicon. Such approaches work only to a limited extent and lead to over-generation of derivational relations, which is a bottleneck of the existing machine-trackable resources of Russian derivational morphology.

In this paper, we have two goals. We present an extension of the novelty methods of constructing language resources for derivational morphology, and we applied this process to a lexicon of Russian. As for the methods, we start with a state-of-the-art grammar-based derivational model which searches for derivational relations between words from the given lexicon (Vodolazsky, 2020). This component works better than regular expressions but still returns many potential derivational relations for each word. Therefore, we re-implement the existing procedure that was originally proposed for harmonising the existing language resources into the same annotation scheme (Kyjánek et al., 2020). Our implementation consists of a supervised machine learning model that classifies the acceptability of the relations with respect to the Russian derivational morphology, and an algorithm for finding maximum spanning trees to select the correct relations within the whole nests of derivationally related words (hereafter *derivational families*) proposed by the grammar-based component. The key difference between the original implementation meant for harmonising and our implementation is that we train the two mentioned components

[1]Reynolds (2016, pp.12-55) provides an overview of resources and tools for processing Russian inflection.

[2]It should be mentioned here that a boundary between the two morphological processes is fuzzy. Linguists discuss many criteria for defining, delimiting, and modelling them as different phenomena, cf. ten Hacken (2014). However, there are also approaches claiming that they both can be treated and modelled in the same way, cf. derivational paradigms (van Marle, 1985; Štekauer, 2014; Bonami and Strnadová, 2019) resembling paradigms often used for inflectional morphology.

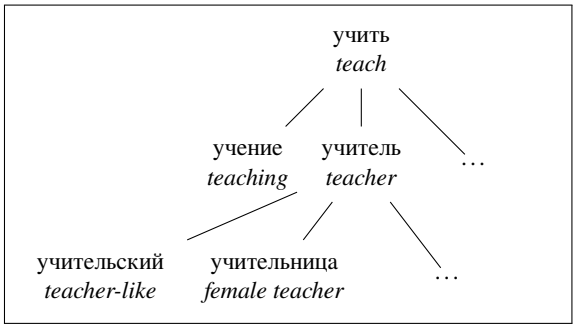

Figure 1: An example of a rooted tree of derivationally related lexemes to the verb учить '*uchit*' (*teach*).

together, while the harmonisation process utilises them as two subsequent steps. As for the application of the described methods to a lexicon extracted from the large web-based Russian corpus Araneum Russicum Maius (Benko, 2014), we construct a new resource dubbed DeriNet.RU, which raise a new leading edge by being the largest, wide-coverage, and contemporary (regarding the Russian vocabulary) lexical resource of Russian derivational morphology. It contains more than 300 thousand corpus-attested lexemes connected with more than 164 thousand binary derivational relations comprising almost 173 thousand derivational families. The resource models derivational morphology according to morphological complexity of lexemes, as it structures the data into rooted trees (in the terminology of graph theory) as illustrated in Figure 1. We provide DeriNet.RU as freely available under an open license.

The paper is structured as follows. Section 2 describes both the existing methods using which lexical resources of derivational morphology are built and the current situation of these lexical resources for Russian. Section 3 focuses on our implementation leading to the creation of the DeriNet.RU database. In Section 4, we present the statistical properties of the database, and a pilot comparison of the created data with the other existing resources not only for Russian because we harmonise DeriNet.RU into the same annotation scheme as other existing resources of derivational morphology. The last Section 5 concludes the paper and considers the future directions of the research.

2. Related Work

2.1. Resources of Derivational Morphology, especially for Russian

There are many resources of derivational morphology for dozens of mostly European languages, cf. list from Kyjánek (2018). They cover the whole range from resources specialised in derivational morphology to additional annotations of derivations incorporated in corpora or resources primarily intended for other phenomena, e.g., WordNets. The resources differ in many aspects, namely the original annotation schemata, numbers and types of covered lexemes and relations, file formats, licenses, etc.; therefore, Kyjánek et al. (2020) have harmonised several of them into the same annotation scheme and published them under open licenses in the Universal Derivations (UDer) collection. Its crucial design property is that each derivative has at most one morphological base, and thus derivational families are represented as rooted tree graphs structured according to the morphological complexity of lexemes; the root represents the shortest lexeme, and morphological complexity gradually grows towards leaf nodes.

In the case of Russian, we know five digital data resources specialised in derivational morphology. Each of them, however, suffers from one or more significant drawbacks, either in vocabulary size and number of relations or in availability.

- **Slovoobrazovatelnyj slovar russkogo jazyka** [*Word-Formation Dictionary of Russian*] (Tikhonov, 1985) is a resource of more than 145 thousand lexemes from which nearly 60 thousand lexemes have been extracted and digitised. The license of this data set is, however, not specified and thus unclear.
- **Slovar morfem russkogo jazyka** [*Dictionary of Morphemes of Russian*] (Kuznetsova and Efremova, 1986) is a resource of around 52 thousand morphologically segmented lexemes. It has been digitised and expanded. It does not contain any explicit derivational relations, but lexemes belonging to the same derivational family can be identified based on the root morphemes which are labelled and resolved for allomorphy. Its license is not specified.
- **Russian Derivational Morphology Database (Unimorph)** (Augerot, 2002) is a data set of almost 93 thousand derivationally related lexemes taken from the Grammar dictionary of Russian (Zaliznyak, 1977). The resource is available online for querying morphemes or strings only.[3]
- **Database DerivBase.Ru** (Vodolazsky, 2020) is an open-source Python library[4] that returns derivatives for a given lexeme. It is based on a set of manually written rules that covers the most productive derivational types in Russian, including derivatives with special alternations of characters. The library has been applied to the lexicon of more than 270 thousand lexemes from Russian Wiktionary and Wikipedia. The resulting database has been additionally harmonised and released in the Universal Derivations collection under the Creative Commons license.
- **Russian DeriNet** (Ignashina, 2020) is a database of Russian derivational relations merged from the first two resources mentioned above.[5] It has been

[3] http://courses.washington.edu/unimorph/

[4] https://github.com/s231644/DerivBaseRu

[5] https://github.com/mashashaitz/Russian-Derinet

harmonised into the Universal Derivations framework but has not been incorporated in the collection because of the unclear licenses of the input resources.

2.2. Methods of Creating Resources for Derivational Morphology

When reviewing ways of creating the existing language resources of derivational morphology, we basically observed either manually or automatically created resources. The resources that are constructed completely manually, e.g., the first three above-mentioned Russian resources and CELEX2 for Dutch, English, and German (Baayen et al., 1995) have very high precision but if these resources are not small, this approach is expensive as it requires many annotators. On the other hand, the automatically created resources must strongly rely on the regularity of derivational morphology, and thus their precision is lower but their size can be bigger.

The process of automatic creation of resources for derivational morphology of any language starts with as large a lexicon of lexemes from a language as possible, and the key task is to search for derivationally related lexemes within this lexicon. As for the above-mentioned Russian resources, DeriNet.Ru has been created in such way. Its author, inspired by the work on German resource DErivBase (Zeller et al., 2013), searched for the derivationally related lexemes by using a rule-based framework, which resembles regular expressions but achieves better results as it can handle complicated alternations of letters during derivational processes. The rules applied to the lexicon were extracted from the existing grammar books of Russian. A similar approach has been also presented by Baranes and Sagot (2014) and Lango et al. (2021). They exploit techniques of pattern matching from the field of machine learning. A machine-learning model extracts the rules on its own on the basis of the data. It is a language-independent approach, but it over-generates the potential derivational relations.

A more advanced way is to create a resource using a tool for morphological segmentation of stemming if there is such tool, and cluster lexemes into flat sets of derivationally related lexemes (cf. Gaussier (1999)) or into structured derivational families (cf. Haghdoost et al. (2019)) according to the segmented units. In addition to the need for such a tool, this approach also suffers from the necessity of distinguishing root morphemes from affixes to be able to cluster lexemes properly.

3. Construction of the New Lexical Resource DeriNet.RU

We create a new lexical resource of Russian derivational morphology that is wide-coverage in terms of its vocabulary and that stores the data in the same way as other existing resources of this kind, i.e., from the Universal Derivations collection. While the first requirement is fulfilled by selecting a sufficiently large Russian corpus, the second one by choosing the data structure of rooted trees for representing derivational families (i.e., lexemes are assigned to nodes, derivational relations are edges between nodes, each node can be connected by at most one antecedent) as illustrated on Figure 1. This tree-based approach takes derivation to be a binary relation, concurring with the linguistic notion by Dokulil (1962) and Körtvélyessy et al. (2020).[6] Since this data structure and the resulting file format originate from the Czech resource of derivational morphology DeriNet (Vidra et al., 2021), we call the new resource of Russian derivational morphology as DeriNet.RU.

As for the process of constructing the resource, we exploit a novel combination of the existing rule-based model of Russian derivational morphology created by Vodolazsky (2020) and the procedure used for harmonising the existing resources into the rooted tree data structure (Kyjánek et al., 2020). However, the latter technique needs re-implementation to be able to utilise it for the construction of a new resource. In order to create and primarily evaluate the resulting resource, we also create a golden data set of manual, parallel annotations of Russian derivational morphology.

In the following subsections, we describe three steps of constructing DeriNet.RU and close this section with a brief discussion of unclear cases of relations from the data we observe during the work. Section 3.1 is devoted to a compilation of an underlying, large set of corpus-attested lexemes. Section 3.2 focuses on searching for derivational relations within the lexicon. Section 3.3 describes our re-implementation of techniques for structuring derivationally related lexemes into rooted trees. Section 3.4 presents our solutions to several types of relations that allow more ways of their modelling in the rooted tree data structure.

3.1. Compiling the Underlying Lexicon

As lexemes are the basic units for the tree-based model of derivations and they also serve as a basis for searching for derivational relations, the compilation of a large underlying lexicon is crucial. We have observed two ways of creating such a set of lexemes for Russian so far. The first one merges the already existing resources (cf. Russian DeriNet). We cannot go this way because the above-presented resources have restrictive licenses which would prevent us from publishing the resulting data resource as open-source. In addition, the lexicon should contain corpus-attested lexemes if it is to be widely used and if the process of searching for derivational relations is to have high recall and precision. The second way is to build a resource based on Russian Wikipedia and Wiktionary (cf. DerivBase.Ru). Such data set contains a disproportionate amount of terminological lexemes compared to more common ones.

[6] There are also graph-based approaches designed in other ways, without the treeness constraint. For example, the data structure of German DErivBase (Zeller et al., 2013), French Démonette (Hathout and Namer, 2014) and other resources allow more antecedents for each lexeme.

We believe this approach is satisfying enough but it is reasonable to consider using a more *representative* text corpus of Russian. Two of the existing Russian corpora can be taken into account:

- **Russian National Corpus**[7] is a representative corpus of modern Russian texts that incorporates almost 6 million lexemes. The corpus is tokenised, lemmatised, and morpho-syntactically analysed. However, its license terms prevent us from using it in the process of compiling any new resource.
- **Aranaeum Russicum Corpus**[8] is the Russian portion of the Aranea project which assembles a family of comparable corpora sharing certain properties with several other languages, e.g., tagset. Araneum Russicum is distributed in two different sizes: Mius and Maius. Both versions are tokenised, lemmatised, and morpho-syntactically analysed. They are hosted by the Czech National Corpus, which provides the KonText interface for querying the corpora.[9]

For license reasons, we have decided to exploit Araneum Russicum Maius. We extract nouns, adjectives, verbs and adverbs with their frequency counts from the corpus. We exclude lexemes that occur less than five times in the corpus to prevent the resulting set of lexemes from containing typos and low-frequency lexemes. Pronouns, numerals and other part-of-speech categories are not included – their derivations are subject to future work.

To give an example of problems we faced during the creation of the set of lexemes, we mention encoding, lemmatisation, and tagging. Some extracted lexemes in Araneum contain Latin character encoding instead of Cyrillic. Especially the visually same characters, such as A, B, E, occurred between Cyrillic ones. We correct their encoding on the basis of the Unicode standard. These standards help us to exclude lexemes belonging to vocabularies of other languages, such as Arabic, Greek, and Hebrew, as well as lexemes containing punctuation, except for a dash which is often used in Russian compounds. As for the atypical tagging (e.g., the noun Шаня '*Shanya*' (*River Shanya*) is tagged as a verb, or the verb полировать '*polirovat*' (*to polish*) is tagged as both noun and adjective) and lemmatisation of adjectives and verbs (the corpus includes some inflectional forms of adjectives and verbs as separate lemmas), we exploit regular expressions and the Russian automatic morphological analyser `pymorphy2` (Korobov, 2015). However, we adhere to the original lemmatisation in the cases of negated lexemes and active participles from the Araneum Russicum Corpus (we keep both the affirmative and negated lexemes as well as the active and passive participles), as it corresponds to the Russian linguistic traditions.

[7] `https://ruscorpora.ru/`

[8] `http://ucts.uniba.sk/aranea_about/_russicum.html`

[9] `https://kontext.korpus.cz/`

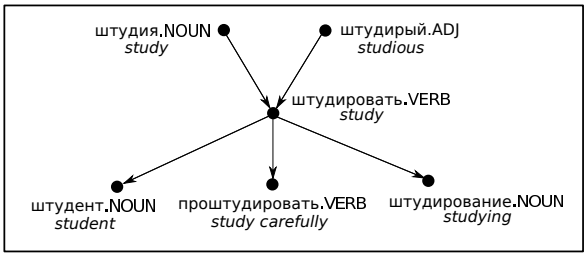

Figure 2: The noun штудировать '*shtudirovat*' (*study*) and its derivational family proposed by a grammar-based component.

Gold data In order to obtain a data set that we can use to develop and evaluate our work, we have sampled 500 lexemes and annotated their derivational relations manually. The lexemes have been sampled randomly from a probabilistic distribution proportionally to absolute counts of occurrences in the Araneum Russicum Corpus. Two Russian native speakers with linguistic education annotated the sample independently in parallel; their task was to determine a base lexeme (i.e., derivational antecedent with less morphological complexity) for each sampled lexeme. Their inter-annotator agreement was 80%; their disagreements in the remaining 20% were resolved by both annotators working together to produce a single solution for each sampled lexeme.

3.2. Searching for Derivational Relations and Building Families

Derivational relations between lexemes are searched for using a grammar-based component. It consists of a manually created set of derivational rules extracted from Shvedova (1980), and it was exploited when DerivBase.Ru was built. Besides the formal structures, including character alternations during derivations, and part-of-speech categories of base and derived lexemes, the rules also consider various inflectional categories analysed by the morphological analyser `pymorphy2`. Some examples of the rules are given below:

- **rule343**(noun + ист → noun), e.g., анархия '*anarkhiya*' (*anarchy*) → анархист '*anarkhist*' (*anarchist*);
- **rule887**(у + adj + и1(ть) → verb), e.g., простой '*prostoj*' (*simple*) → упростить '*uprostit*' (*to simplify*).

When applying the component, we start with the set of lexemes. For each base lexeme l_b, we add a relation to those potentially derivative lexemes l_d which, according to the system of derivational rules, could be derived from l_b with a rule r. We obtain 1,256,222 candidates for derivational relations by applying the rules to our set of lexemes. If these relations and lexemes are connected, we obtain a model of derivational families that allows more than one base lexeme for a derivative; see Figure 2 which shows an example of one such family.

Table 1: Oracle scores on the gold data using different sets of candidates for derivational relations.

Method of generating candidates for relations	Oracle	Number of relations
(A) complete graph from the set of lexemes	99.8	379,677,792,400
(E) complete graphs from grammar-based c. & interval $n = 15$	95.7	10,406,682,343
(E) complete graphs from grammar-based c. & interval $n = 5$	94.9	10,394,358,943
(D) grammar-based c. & interval approach with $n = 15$	89.9	19,125,212
(D) grammar-based c. & interval approach with $n = 5$	88.5	6,801,812
(B) **grammar-based component**	**87.3**	**1,256,222**
(C) interval approach with $n = 15$	57.7	18,485,175
(C) interval approach with $n = 5$	47.8	6,161,775

To measure how successful this approach is, we define and calculate an *oracle score* on our gold annotations. This score represents a possible maximum of reachable derivational relations taken from the gold data and found in generated candidates for relations. In other words, the oracle score is the maximum accuracy achievable on the gold data when the given set of relations is used. Table 1 presents oracle scores and numbers of relations generated when five different approaches for searching derivational relations are exploited:

(A) If we generated a set of all possible binary relations between all lexemes from our set of lexemes (resulting in one large complete graph), the achievable maximum would be 99% but the number of relations would be too high to predict and to identify rooted trees.

(B) The oracle score for a set of relations generated using the grammar-based component is lower (87.3%) but the number of relations has decreased significantly.

(C) We also tested how the score changes if candidates for the relations are created using the interval of size n lexemes (i.e., n lexemes left and n right) for each lexeme in the lexicographically ordered set of lexemes.[10] This approach did not yield better oracle scores (48-58%); however, it does seem to be useful when no linguistic knowledge is available.

(D) A combination of the interval approach and the grammar-based component increases oracle scores (89-90%), as well as the number of relations.

(E) Complete graphs created from derivationally related lexemes resulted from the interval-based approach and the grammar-based component increase oracle scores noticeably (95-96%) but the number of relations increases rapidly.

[10]Given the definition of derivation, we assume that derivationally related lexemes are lexicographically closer, especially in the case of suffixation. In the case of prefixation, the characters of lexemes could be changed to retrograde order, and the same interval-based approach would look for prefixed lexemes.

As a result of this analysis, we used candidates for derivational relations proposed by the grammar-based component only, so our maximum achievable accuracy is 87.3%.

3.3. Restructuring Derivational Families into Rooted Trees

Having derivational families with over-generated derivational relations as illustrated in Figure 2, we now want to organise families into rooted trees. The process of identifying rooted trees in the given (weakly connected) graphs exploits a combination of:

- a scorer, i.e. a supervised machine learning model that classifies the acceptability of a relation (by giving a score ranging from 0 to 1 to the classified relation, meaning whether the relations should be present or absent in the resulting rooted tree according to derivational morphology of Russian) and
- a Maximum Spanning Tree algorithm (MST) that identifies the desired (most probable) rooted trees as a maximum sum of scores. The scores are used as weights of edges, i.e., derivational relations. Specifically, we use Chu-Liu/Edmond's algorithm (Chu and Liu, 1965; Edmonds, 1967).

Kyjánek et al. (2020) have proposed a similar combination of the two components when they harmonised the existing resources of derivational morphology into the Universal Derivation collection. We modify this combination to be able to create a new resource: we train and evaluate the two components together as one component which, consequently, allows us to achieve a higher accuracy because the scores given by the scorer can reflect trees resulting from the MST algorithm.

Development We cross-validate the model using 10-folds from our gold data. As the gold data contains only 500 positive examples, i.e., derivational relations that should be present in the resulting trees, we create another 500 negative examples automatically. All one thousand relations are assigned with the following features: part-of-speech categories of a base lexeme and its derivative, Levenshtein distance (Levenshtein, 1966), Jaro-Winkler distance (Jaro, 1989; Winkler, 1990), Jaccard distance (Jaccard, 1912), length of

Table 2: Accuracy of cross-validated different methods and a baseline model for identification of rooted trees. The machine learning methods implemented in `scikit-learn v0.23.2` are exploited as the scorer. Their hyper-parameters are specified in the table, otherwise, default values remain. Chu-Liu/Edmond's algorithm is used for finding Maximum Spanning Trees.

Method of scoring candidates for relations (+MST)	**Accuracy**
RandomForestClassifier(criterion=entropy, max_depth=5, n_estimators=400, min_impurity_decrease=0.01)	**62.9**
AdaBoostClassifier(n_estimators=300, learning_rate=0.001)	59.9
BernoulliNB()	59.5
CalibratedClassifierCV(Perceptron(max_iter=2000, penalty=l1, alpha=0.01), cv=20)	59.5
MLPClassifier(hidden_layer_sizes=(100,)*50, activation=logistic, max_iter=1000)	59.1
DecisionTreeClassifier(criterion=entropy, min_impurity_decrease=0.0001)	59.0
LogisticRegression(solver=newton-cg, multi_class=ovr, max_iter=5000)	55.0
KNeighborsClassifier(n_neighbors=5)	54.2
Baseline(random score)	**52.6**

the longest common substring, boolean values of (i) being base lexeme shorter than its derivative and (ii) having the same initial and final character n-grams (for $n = \{1, 2\}$), percentage intersections of (i) all characters and (ii) consonants of a base lexeme and its derivative, and string forms of the initial and final character n-grams (for $n = \{1, 2, 3, 4, 5\}$) of a base lexeme and its derivative. The categorical values are one-hot encoded; the boolean and numeric values remain the same.

Evaluation When evaluating the models, we extract the entire derivational families (from the data set of all derivational relations proposed by the grammar-based component) from the testing folds. These relations are predicted and their scores serve as weights of graph edges during the identification of rooted trees using the MST algorithm. The accuracy is calculated as the metric of how many relations are identified correctly.

Selection of the best combination Several machine learning methods are tested for the task: Naive Bayes, K-Nearest Neighbour, Decision Trees, Logistic Regression, Random Forest, AdaBoost, Perceptron, and Multilayer Perceptron. Their hyper-parameters are tuned using Grid search. Table 2 shows the best accuracy obtained by each method, as well as the accuracy of the Baseline model (52.6%) that predicted scores randomly in the range $[0; 1]$. The Random Forest method achieved the best results (62.9%).

Application Using the best setting, we applied it to the set of candidates for the derivational relations resulting from the grammar-based component. Figure 3 illustrates one derivational family organised in a rooted tree. The basic quantitative properties of the resulting DeriNet.RU are presented in Table 3 in comparison with other existing resources for Russian derivation morphology.

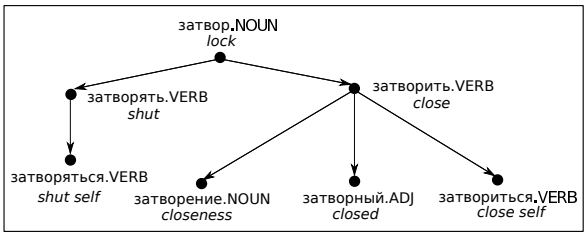

Figure 3: Derivational family of the lexeme затвор '*zatvor*' (*lock*) represented as a rooted tree.

3.4. Fuzzy Cases in the Russian Derivational Morphology

During the manual annotations and the work on DeriNet.RU, we have noticed some cases whose modelling are fuzzy as they refer to a potential borderline between inflectional and derivational morphology, although the borderline is far from self-evident. In practice, some decisions about these cases are arbitrary, different from language to language or specific to a particular linguistic tradition. The approaches by Shvedova (1980), Plungian (2003), Vinogradov (1972), Shansky and Tikhonov (1987), and Pertsov (2001) present the expected divergences and the consequent (different) decisions in the Russian linguistic tradition. We show two points of divergence that illustrate different decisions when it comes to delimiting the boundary between inflection and derivation in Russian and their consequences for modelling derivations.

Representation of lexical (affixal) negation, e.g. правильный '*pravilnyy*' (*correct*) and неправильный '*nepravilnyy*' (*incorrect*), in Russian is considered to be derivation because of the preference for the criterion of semantic regularity in the Russian lexicographic tradition. Most lexemes in Russian can be negated, but the meaning of the base and derived lexemes can often diverge significantly, e.g., посредственный '*posredstvennyj*' (*mediocre*) vs. непосредственный '*neposredstvennyj*' (*direct*). However, in a related

Table 3: Quantitative properties of the existing resources compared to the resulting DeriNet.RU. Only nouns, adjectives, verbs and adverbs have been extracted from Araneum Russicum Maius. Columns *Tree size*, *Tree depth*, and *Tree out-degree* are presented in average / maximum value format. Part-of-speech distribution is ordered as follows: nouns, adjectives, verbs, adverbs, and other categories.

Resource	Lex	Rel	Fam	Singl	Size	Depth	Out-deg	POS distr.
Slov. slovar (Tikhonov, 1985)	59,265	56,633	2,632	1,019	22.5/504	1.7/8	7.8/280	42/27/27/4/1
DerivBase.Ru (Vodolazsky, 2020)	270,473	133,759	136,714	116,036	2.0/1142	0.3/13	0.4/36	62/18/17/3/0
Russian DeriNet (Ignashina, 2020)	88,180	66,184	21,996	15,113	4.0/484	0.6/10	1.3/238	41/28/28/3/0
DeriNet.RU	**337,632**	**164,725**	**172,907**	**99,624**	**2.0/586**	**0.6/14**	**0.6/24**	**58/19/20/3/0**

Slavic language, in Czech, lexical negation is analysed as inflection, although it behaves the same because the Czech tradition prioritises syntactic regularity. As the result, both negated and affirmative lexemes are represented by separate lemmas in Russian language resources, while the resources for Czech represent the negated and affirmative lexemes by one same lemma. Having two separate negated and affirmative lexemes then raise the question of whether to model these negated relations apart from affirmative relations (i.e., as two parallel sub-trees), or whether keeping the negated lexemes as a direct descendant of their affirmative lexemes. We keep both negated and affirmative lexemes and model their relations in the latter way.

The opposite situation is in the case of active participles, e.g., одевать '*odevat*' (*to dress*) and одевающий '*odevayuschij*' (*dressing*). For Russian, participles are generally considered to belong to a particular verbal inflectional paradigm (Lyashevskaya et al., 2005) and are lemmatised together with the corresponding verb. However, the active participles are missing when modelling their descendants, and thus we represent the active and passive participles under different lemmas.

4. Evaluation

Having a gold sample consisting of 500 correct and 500 incorrect derivational relations, we calculated an oracle score, i.e. the maximum accuracy achievable on the basis of a given set of candidates for derivational relations (87.3%). Furthermore, the gold data were used for identifying and evaluating the rooted trees. While the baseline model achieved an accuracy of 52.6%, the best supervised machine-learning model that uses Random Forest achieved 62.9%.

In the following subsections, we present a comparison of the resulting DeriNet.RU with other existing resources of derivational morphology for Russian (from Section 2.1) and for other languages incorporated in the Universal Derivations collection, i.e., the collection whose data structure and file format we adopted for DeriNet.RU. While the former comparison serves as an indirect evaluation of data quality, the latter one should serve as a pilot cross-linguistic overview.

4.1. Comparison to Russian Resources

Table 3 summarises some statistic properties of the data sets. DeriNet.RU outperforms other existing resources in almost all presented aspects. It contains the most lexemes, derivational relations and families; moreover, the lexemes are corpus-attested. The number of so-called singletons (lexemes that are not connected to any other lexemes) is relatively small, in comparison to other resources. DeriNet.RU contains smaller derivational families than other resources, but their depth and out-degree[11] are comparable to other resources. This trend is observable also in other (semi-)automatically created resources, namely DerivBase.Ru and Russian DeriNet. The part-of-speech category distributions are comparable across all resources.

4.2. Comparison to UDer Resources

Figure 4 shows a comparison of the 15 resources from the Universal Derivations collection (including DeriNet.RU) that capture the most lexemes. DeriNet.RU is the second biggest resource regarding the number of lexemes, and it is between the top three resources in terms of captured derivational relations and families. Its number of singletons, i.e., derivational families consisting of only one lexeme, is comparable to the number of singletons in Czech DeriNet, which is created semi-automatically with a high precision.

As for the distribution of part-of-speech categories in the harmonised resources, only 11 resources are tagged. Their distributions of part-of-speech categories are relatively similar, except for German DerivBase, which contain significantly more nouns than other categories. To give a better insight into the structures of derivational families, we measured their average sizes, depths (i.e., how many subsequent derivations are in a family) and out-degrees (i.e., how many lexemes can be derived from a single lexeme). Compared to other resources, DeriNet.RU contains medium-sizes families whose depths and out-degrees are relatively low, but still much bigger than DerivBase.Ru.

When analysing distributions of derivational relations

[11] Tree out-degree, here, is the maximum number of nodes to which the node in consideration points in a tree (derivatives derived from base lexemes). It simply represents the width of a tree.

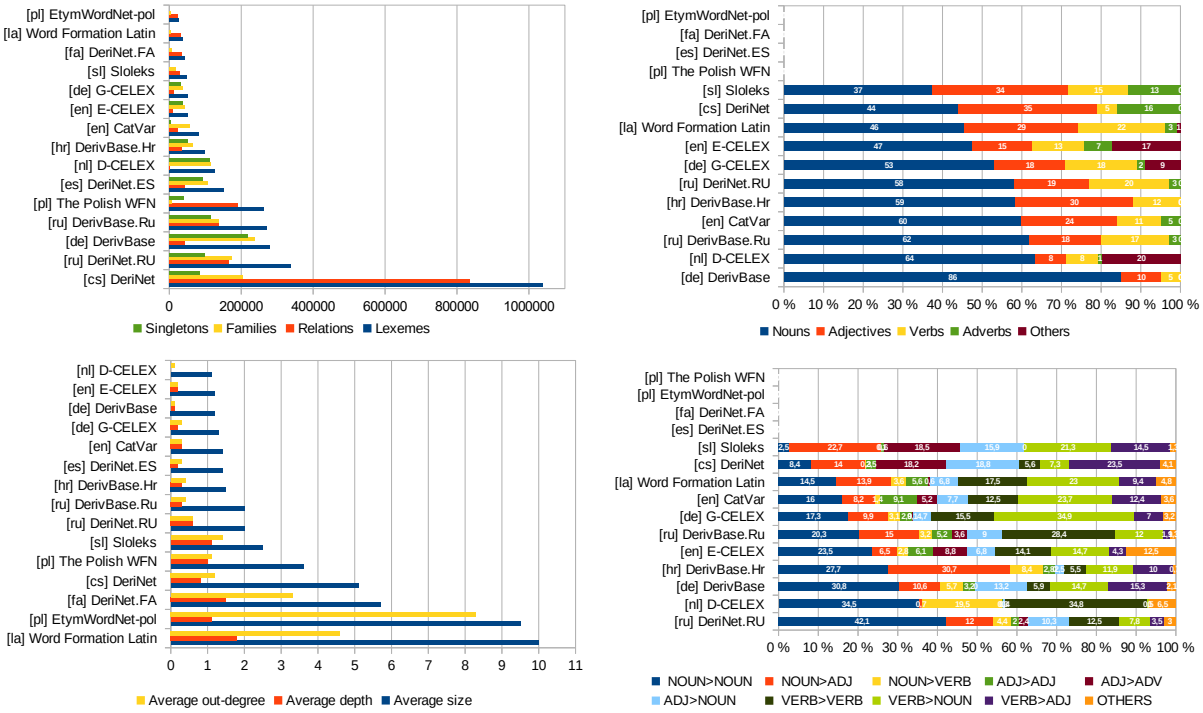

Figure 4: Quantitative comparison between the top 15 UDer resources and DeriNet.RU.

according to the part-of-speech categories of a base lexeme and its derivative, we can see that DeriNet.RU includes 42.1 derivational relations in which both a base lexeme and its derivative are nouns; it is the most of all the resources. The proportion of relations in DeriNet.RU seems similar to DerivBase.Ru, which is due to the usage of the same grammar-based component.

In conclusion, the above-presented comparison of DeriNet.RU and other resources from Universal Derivations shows that the new resource we constructed can withstand comparisons with the largest existing resources of derivational morphology across languages. Although the individual resources stand for different languages, the presented numbers should not be interpreted as a cross-linguistic comparison. However, it still shows a strong potential for making such interpretations on the basis of the harmonised resources.

5. Conclusions

This paper has presented the way of constructing the largest lexical resource of Russian derivational morphology and the resulting resource dubbed DeriNet.RU. It is ready for use in the NLP tasks as well as a data background for linguistic research, which we illustrate in the comparison of DeriNet.RU with other resources harmonised in the same annotation scheme (from the Universal Derivations collection). The resource has been created based on the state-of-the-art methods (using a combination of a grammar-based component, a supervised machine learning scorer and a Maximum Spanning Tree algorithm), and it contains a large, wide-coverage and corpus-attested set of lexemes from the contemporary Russian language. DeriNet.RU is released in the Universal Derivations collection v1.1 under the Creative Commons license (CC BY-NC-SA 4.0) and is available for querying via DeriSearch.[12]

The crucial part of this paper is the construction of a new resource. We believe that the presented process and the discussed methods are as replicable as possible. Besides, we addressed some issues which are relevant to all similar data resources but are hardly ever discussed, e.g., the way of compiling the underlying set of lexemes of the resulting resource, fuzzy cases of modelling morphology, such as lexical affixal negation, and the choice of an appropriate evaluation. In this work, we limited the resource to include only derivations within and between nouns, adjectives, verbs, and adverbs, but not other part-of-speech categories or compounding. We are leaving these in more detail for our future work.

6. Acknowledgements

This work was supported by the Grant No. GA19-14534S of the Czech Science Foundation, the Grant No. START/HUM/010 of Grant schemes at Charles University (reg. No. CZ.02.2.69/0.0/0.0/19_073/0016935), the LINDAT/CLARIAH-CZ project of the Ministry of Education (LM2015071, LM2018101), Youth and Sports of the Czech Republic (project LM2018101), and by the SVV project No. 260 575. It was using language resources developed, stored, and distributed by the LINDAT/CLARIAH-CZ project. We thank Novosibirsk State University for the computational time on HCI NSU graphic cluster.

[12] http://ufal.cz/derisearch

7. Bibliographical References

Baranes, M. and Sagot, B. (2014). A Language-independent Approach to Extracting Derivational Relations from an Inflectional Lexicon. In *LREC*, volume 14, pages 2793–2799. Citeseer.

Bonami, O. and Strnadová, J. (2019). Paradigm Structure and Predictability in Derivational Morphology. *Morphology*, 29:167–197.

Chu, Y.-J. and Liu, T. H. (1965). On the Shortest Arborescence of a Directed Graph. *Scientia Sinica*, 14:1396–1400.

Dokulil, M. (1962). *Tvoření slov v češtině 1: Teorie odvozování slov*. Academia, Prague.

Edmonds, J. (1967). Optimum Branchings. *Journal of Research of the national Bureau of Standards*, 71B(4):233–240.

Gaussier, É. (1999). Unsupervised Learning of Derivational Morphology from Inflectional Lexicons. In *Unsupervised Learning in Natural Language Processing*.

Haghdoost, H., Ansari, E., Žabokrtský, Z., and Nikravesh, M. (2019). Building a Morphological Network for Persian on Top of a Morpheme-Segmented Lexicon. In *Proceedings of the 2nd Workshop on Resources and Tools for Derivational Morphology*. Charles University.

Jaccard, P. (1912). The Distribution of the Flora in the Alpine Zone. *New Phytologist*, 11(2):37–50.

Jaro, M. A. (1989). Advances in Record-linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association*, 84(406):414–420.

Korobov, M. (2015). Morphological Analyzer and Generator for Russian and Ukrainian Languages. In *Analysis of Images, Social Networks and Texts*, volume 542 of *Communications in Computer and Information Science*, pages 320–332. Springer International Publishing.

Lívia Körtvélyessy, et al., editors. (2020). *Derivational Networks Across Languages*. De Gruyter Mouton.

Kyjánek, L., Žabokrtský, Z., Ševčíková, M., and Vidra, J. (2020). Universal Derivations 1.0, A Growing Collection of Harmonised Word-Formation Resources. *The Prague Bulletin of Mathematical Linguistics*, 115:5–30, October.

Kyjánek, L. (2018). Morphological Resources of Derivational Word-Formation Relations. Technical Report TR-2018-61, Faculty of Mathematics and Physics, Charles University.

Körtvélyessy, L. (2016). Word-formation in Slavic languages. *Poznań Studies in Contemporary Linguistics*, pages 455–501.

Lango, M., Žabokrtský, Z., and Ševčíková, M. (2021). Semi-automatic Construction of Word-formation Networks. *Language Resources and Evaluation*, 55:3–32.

Levenshtein, V. I. (1966). Binary Codes Capable of Correcting Deletions, Insertions and Reversals. *Soviet Physics Doklady*, 10(8):707–710.

Lipka, L. (2010). *An Outline of English Lexicology*. De Gruyter.

Lyashevskaya, O., Plungian, V., and Sitchinava, D. (2005). O morphologicheskom standarte Nacional'nogo korpusa russkogo jazyka [About the morphological standard of the Russian National Corpus]. In *Nacional'nyj korpus russkogo jazyka: 2003-2005. Rezul'taty i perspektivy*, pages 111–135. Rossijskaja Akad. Nauk.

Pertsov, N. V. (2001). *Invarianty v russkom slovoizmenenii [Invariants in Russian inflection]*. Jazyki Slavjaskoj Kul'tury, Moscow.

Plungian, V. (2003). *Obschaja morfologija: Vvedenije v problematiku [General morphology: An introduction]*. URSS.

Reynolds, R. (2016). *Russian Natural Language Processing for Computer-assisted Language Learning: Capturing the Benefits of Deep Morphological Analysis in Real-life Applications*. Ph.D. thesis, UiT The Arctic University of Norway, Faculty of Humanities, Social Sciences, and Education.

Shansky, N. M. and Tikhonov, A. N. (1987). *Sovremennyj russkij jazyk [Modern Russian Language]*. Prosveschenije, Moscow.

Shvedova, N. (1980). *Russkaja grammatika [Russian grammar]*. Number t. 1 in Russkaja grammatika. Izd-vo Nauka.

Štekauer, P. (2014). Derivational Paradigms. In Rochelle Lieber et al., editors, *The Oxford Handbook of Derivational Morphology*, pages 354–369. Oxford University Press, Oxford.

Štekauer, P., Valera, S., and Körtvélyessy, L. (2012). *Word-Formation in the World's Languages: A Typological Survey*. Cambridge University Press, New York.

ten Hacken, P. (2014). Delineating Derivation and Inflection. In Rochelle Lieber et al., editors, *The Oxford Handbook of Derivational Morphology*, pages 10–25. Oxford University Press, Oxford.

van Marle, J. (1985). *On the Paradigmatic Dimension of Morphological Creativity*. Walter de Gruyter GmbH & Co KG, Dordrecht.

Vinogradov, V. V. (1972). *Russkij jazyk [The Russian language]*. Vysšaja škola, Moscow.

Winkler, W. E. (1990). String Comparator Metrics and Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. In *Proceedings of the Section on Survey Research*, pages 354–359. ERIC.

8. Language Resource References

Augerot, J. (2002). Russian Morphological Database.

Baayen, H. R., Piepenbrock, R., and Gulikers, L. (1995). CELEX2. Linguistic Data Consortium, Catalogue No. LDC96L14.

Benko, V. (2014). Aranea: Yet Another Family of (Comparable) Web Corpora. In *Text, Speech, and Di-*

alogue International Conference (TSD), pages 247–256. Springer.

Hathout, N. and Namer, F. (2014). Démonette, a French Derivational Morpho-Semantic Network. *Linguistic Issues in Language Technology*, 11:125–162.

Ignashina, M. (2020). Russian DeriNet: from Word Formation Dictionaries to Universal Derivations. Master's thesis, National Research University Higher School of Economics, Faculty of Humanities.

Kuznetsova, A. I. and Efremova, T. F. (1986). *Slovar' morfem russkogo jazyka [Dictionary of morphemes of the Russian language]*. Russkij jazyk, Moscow.

Tikhonov, A. N. (1985). *Slovoobrazovatel'nyj slovar' russkogo jazyka: v 2 t. Ok. 145 000 slov [Word-formation dictionary of Russian: in 2 volumes. Aprx. 145000 words]*. Russkij jazyk, Moscow.

Vidra, J., Žabokrtský, Z., Kyjánek, L., Ševčíková, M., Dohnalová, Š., Svoboda, E., and Bodnár, J. (2021). DeriNet 2.1. LINDAT/CLARIAH-CZ, ÚFAL, Faculty of Mathematics and Physics, Charles University.

Vodolazsky, D. (2020). DerivBase.Ru: A Derivational Morphology Resource for Russian. In *LREC*, volume 20, pages 3930–3936, Marseille, France.

Zaliznyak, A. A. (1977). *Grammaticheskij slovar russkogo jazyka. Slovoizmenenije. [Grammar dictionary of the Russian language. Inflection.]*. Russkij jazyk.

Zeller, B., Šnajder, J., and Padó, S. (2013). DErivBase: Inducing and Evaluating a Derivational Morphology Resource for German. In *ACL*, volume 1, pages 1201–1211.