DeriNet 2.0: Towards an All-in-One Word-Formation Resource

Jonáš Vidra **Zdeněk Žabokrtský** **Magda Ševčíková** **Lukáš Kyjánek**
Charles University, Faculty of Mathematics and Physics,
Institute of Formal and Applied Linguistics
Malostranské náměstí 25, 118 00 Prague 1, Czech Republic
{vidra,zabokrtsky,sevcikova,kyjanek}@ufal.mff.cuni.cz

Abstract

DeriNet is a large linguistic resource containing over 1 million lexemes of Czech connected by almost 810 thousand links that correspond to derivational relations. In the previous version, DeriNet 1.7, it only contained very sparse annotations of features other than derivations – it listed the lemma and part-of-speech category of each lexeme and since version 1.5, a true/false flag with lexemes created by compounding.

The paper presents an extended version of this network, labelled DeriNet 2.0, which adds a number of features, namely annotation of morphological categories (aspect, gender and animacy) with all lexemes in the database, identification of root morphemes in 250 thousand lexemes, annotation of five semantic labels (diminutive, possessive, female, iterative, and aspect) with 150 thousand derivational relations, a pilot annotation of parents of compounds, and another pilot annotation of so-called fictitious lexemes, which connect related derivational families without a common synchronous parent. The new pieces of annotation could be added thanks to a new file format for storing the network, which aims to be general and extensible, and therefore possibly usable to other similar projects.

1 Motivation

The paper deals with extending DeriNet, a lexical database developed for Czech, which contains around 1 million lexemes connected with app. 810 thousand edges representing morphological derivations (Ševčíková and Žabokrtský, 2014), forming app. 220 thousand tree-shaped derivational families. The resulting version is labelled DeriNet 2.0 (Vidra et al., 2019) and it is available for download under a free non-commercial license. The extension is mostly qualitative: we extended the expressive power of the underlying data structure (and of the associated file format) substantially and thus enabled capturing language phenomena which were impossible to handle in the previous versions of DeriNet. More specifically, there are five newly supported annotation components in the DeriNet annotation scheme:

- **morphological categories**: lexemes are assigned morphological categories that remain constant under inflection, such as gender with nouns or aspect with verbs,
- **morpheme segmentation**: lexemes belonging to the largest derivational families have their root morphemes identified,
- **semantic labels**: derivational relations are assigned labels capturing the change that the meaning of the base word undergoes by attaching the affix (in affixation),
- **compounds**: lexemes with two (or even more) roots are linked with their both (or more) base words. The linking of compounds with their base words has not been possible so far due to the highly constrained data structure used in DeriNet 1.7 and older versions,

Proceedings of the 2nd Int. Workshop on Resources and Tools for Derivational Morphology (DeriMo 2019), pages 81–89,
Prague, Czechia, 19-20 September 2019.

- **fictitious lexemes**: lexemes that are attested neither in the corpora nor in the dictionaries but, based on structural analogies, fill a paradigm gap in the derivational family are newly added into the database.

Feature	1.7	2.0
Derivational relations	✓	✓
Part-of-speech category	✓	✓
Morphological categories	✗	✓
Compounding relations	✗[a]	✓
Semantic labels	✗	✓
Morpheme segmentation	✗	✓[b]
Fictitious lexemes	✗	✓

[a] A yes/no flag marking compounds was encoded in the POS category.
[b] In the present version, only root morphs of a subset of lexemes are annotated. The format allows for marking affixes and allomorph resolution as well, but these annotations are not currently available.

Table 1: Comparison of features available in DeriNet 1.7 and 2.0.

The annotations present in DeriNet 2.0 are compared to the previous versions in Table 1.

The actual recall of the newly added annotations is rather limited, but even the incomplete annotations serve as a proof of concept and show the viability of the new annotation scheme. However, the main ambition of our efforts does not lie in adding several new annotation components, but it is more strategical: in the long term we attempt to accumulate virtually all information related to word-formation in a single data resource (similarly to various kind of syntactic and semantic phenomena being annotated), and thus hopefully profit from new synergies due to combining different possible perspectives on word-formation.

Some of the features are already available in existing data resources, so from this viewpoint DeriNet 2.0 is rather eclectic. For instance, detailed information on morphological categories of lexemes is captured in MorfFlex CZ (Hajič and Hlaváčová, 2013), morpheme segmentation is available in the MorphoChallenge dataset (Kurimo et al., 2009), semantic labels of derivations can be found in Démonette (Hathout and Namer, 2014), compounds are identified in CELEX (Baayen et al., 1995), and fictitious lexemes are introduced in Word Formation Latin (Litta Modignani Picozzi et al., 2016). However, none of these resources, to the best of our knowledge, integrate all the features in one data set.

In addition, we believe that the extended annotation scheme is flexible enough to be sustainable for a longer period of time without major changes. At the same time, we plan to apply the scheme to dozens of other languages, so the scheme is designed to be as language agnostic as possible.

2 New features

2.1 Morphological categories

Lexemes were provided with selected morphological categories in DeriNet 2.0, namely with the category of gender and animacy (with nouns) and the category of grammatical aspect (with verbs), in addition to the part-of-speech category already available in the previous versions of the data. These categories do not change in inflection, and are characteristics associated with lexemes as wholes.

The morphological categories to assign were extracted from the MorfFlex CZ dictionary (Hajič and Hlaváčová, 2013), which enumerates all possible word forms and positional part-of-speech tags for each lexeme. The set of part-of-speech tags of a particular lexeme was merged into a single string, tentatively called a *tag mask*, by comparing individual positions of the different tags. If all tags of the lexeme share the same value at a position, it is copied to the tag mask, otherwise it is replaced by the question mark ("?"). For example, exploiting the part-of-speech tags assigned to the individual forms of the noun *chata* 'cottage' (15 tags in total, including e.g. "NNFS1-----A----", "NNFP3-----A----" or "NNFP7-----A---6"), the tag mask "NNF??-----A---?" was compiled, which encodes that the lexeme is a noun (NN), feminine

	Unique combinations	Lexemes
Lemma	2,599	5,342
Lemma + POS category	2,137	4,353
Lemma + POS category + morph. features	518	1,039

Table 2: Counts of homonymous combinations of various lexeme features with the counts of affected lexemes. By definition, the number of lexemes must be at least twice the number of homonymous combinations, since a feature combination that is not shared by at least two lexemes is not a homonym. The number of lexemes is slightly larger, because some lemmas are shared by up to four lexemes: e.g. *stát*, which can mean either 'a country', 'to stand', 'to stop' or 'to melt down'.

gender (F), affirmative polarity (A). The categories associated with the other positions either vary (cf. the question marks in the positions associated with the categories of number, case, and register), or are not applicable to Czech nouns (such as tense, cf. the positions with '-').

In addition to the tag mask format, the morphological categories listed above were extracted from the masks and stored in DeriNet 2.0 using the Universal Features annotation scheme (Nivre et al., 2016).

This approach to extracting the morphological categories has a very high precision: we were unable to find any errors in the grammatical category of gender in an uniformly randomly selected sample of 100 nouns, and we found two errors in the category of aspect in a sample of 100 verbs.

The recall of the annotation is also high, with 99.6% nouns being assigned a gender category and 93.2% of verbs being assigned an aspect category. The nouns with missing gender annotation are mostly foreign words with unclear or varying gender (such as *image* 'image', which can be masculine inanimate, feminine or neuter, depending on the speaker's preference) and words which can be used to denote both male and female persons (such as *šereda* 'ugly (person), gorgon'). The dictionary we use as the lexeme source, MorfFlex CZ, usually handles these cases by having a separate lexeme for each gender (such that all forms of any one lexeme have identical gender), but some lexemes have forms with different genders, resulting in missing gender annotation after extraction. Verbs with missing aspect annotation are mostly missing the aspect category in the source dictionary, but some (about one in six) are marked as biaspectual – we chose to exclude the annotation of these for the time being due to low precision of this part of the annotation.

The morphological categories can in some cases also be used to distinguish homonymous lexemes. Just as there are pairs of lexemes with identical lemmas, but different part-of-speech categories, there are also pairs of lexemes with identical lemmas and part-of-speech categories, but with a different aspect or gender. Using tag masks combined with lemmas, we are able to uniquely identify 3,314 out of 4,353 lexemes with homonymous lemma-POS combinations in DeriNet 2.0; see Table 2 for detailed counts. Therefore, the tag masks serve as auto-generated readable identifiers (distinguishing e.g. masculine inanimate *mol#NNI??-----A---?* 'mole (unit)' and masculine animate *mol#NNM??-----A---?* 'mill moth'), as opposed to e.g. using opaque numerical indices ('mol#1' and 'mol#2') or manually created descriptions ('mol#grammolecule' and 'mol#butterfly') to distinguish homonyms, which are the methods used by the underlying MorfFlex CZ dictionary.

The homonymous lexemes may or may not be parts of the same derivational family. For instance, the noun *růst* 'growth' and the verb *růst* 'to grow', distinguished by the part-of-speech category, are related derivationally, the former one being converted from the latter one. Compared to that, the noun *tulení* 'hugging' and the adjective *tulení* 'seal' are identical in spelling due to truly random coincidence; they belong to different derivational families (with the root lexemes *tulit (se)* 'to hug' and *tuleň* 'seal', respectively).

Morphological categories captured by the tag masks have been exploited also within the semantic labelling task (Section 2.3).

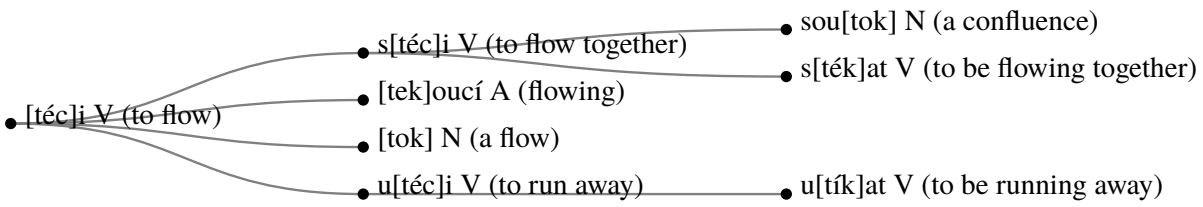

Figure 1: An excerpt from the derivation family of *téci* 'to flow' in DeriNet 2.0, with root morphemes marked by square brackets. Other morphemes are not delimited yet.

2.2 Morpheme segmentation and allomorphy

In DeriNet 2.0, root morphemes of selected lexemes were identified as another new type of annotation. This annotation is currently limited to approx. 250 thousand lexemes and it is supposed to be a sort of pilot approach for a large-coverage morpheme segmentation in the next versions of the data. See Figure 1 for a small sample of the annotation.

Morpheme segmentation, i.e. the task of dividing a word into a sequence of segments corresponding to morphemes as the smallest meaning-bearing language units, is extremely challenging when dealing with Czech. The main reason is the frequent allomorphy of roots and affixes. For instance, in the lexemes that are derivationally related with the verb *jíst* 'to eat' in our data, eight root allomorphs are attested (*jís*, *jíd*, *jed*, *níd*, *nís*, *něd*, *jez*, and *něz*). Notice that there is not a single grapheme shared by all of the allomorphs.

In our first experiment, which aimed at identification of all morphemes in the lexeme structure, we implemented a lemma decision-tree-based segmenter that employed letter n-gram features and was trained using a set of 750 hand-segmented lexemes sampled uniformly randomly from DeriNet. However, the evaluation on an independent dataset showed that the precision of predicted segmentations (95% of identified morphs were correct, resulting in only 85% words being segmented correctly) is below the quality standards usually applied on released versions of DeriNet.[1]

In our second experiment we thus limited the problem to identification of root morphemes and made more intensive use of existing derivational trees. For the 760 biggest trees (in terms of number of nodes), we applied the previously trained segmenter on all lexemes in these trees and tried to distinguish the substring corresponding to the root morpheme in each lexeme using a simple heuristics: for each word, mark its rarest morpheme (measured by the number of occurrences in the whole dataset) as the root; break ties by marking the longer or first such morpheme. We obtained a set of allomorphs of the root morpheme for each tree. The quality of such allomorph sets was relatively low, so the sets were cleaned manually. Then we identified the position of a root allomorph in each lexeme. In case there were multiple matching allomorphs, we preferred the longest one. This process was iterated several times, as applying the allomorph sets to the whole derivational trees uncovered several errors in the annotation of derivational relations. Finally, we added such detected root morpheme boundaries into DeriNet 2.0, which resulted in 243,793 lexemes with identified boundaries of their root morphemes.

There was an interesting side effect of the allomorphy annotations. Some sets of allomorphs for different derivational trees were surprisingly similar. In some cases the string similarity was only due to a random coincidence of etymologically unrelated clusters (such as the derivational family of *řídký* 'sparse' with root allomorphs *řid*, *říd*, *řed* and *řeď*, from which three allomorphs overlap with the family of *řídit* 'to direct, to drive' with allomorphs *řid*, *říz*, *řed*, *řiz* and *říd*), or due to a diachronic etymological relation (since DeriNet focuses on synchronic view of the language, diachronic relations which are opacified in modern language are not included; e.g. *medvěd* 'a bear', which is etymologically a compound with bases *med* 'honey' and *jíst* 'to eat', is not connected to any parents in DeriNet) but sometimes we really revealed a missing relation in DeriNet 1.7; such relations were added into DeriNet 2.0.

[1]One of our design decisions is that when adding new pieces of information into DeriNet, we prefer precision to recall.

Label	Count
POSSESSIVE	88,718
FEMALE	29,023
ASPECT	15,439
ITERATIVE	11,886
DIMINUTIVE	5,939

Table 3: Counts of the semantic labels in DeriNet 2.0 data.

2.3 Semantic labels

Semantic labels, which capture the change in the meaning of the base word imposed by affixation, were assigned with relations in DeriNet as another new type of annotation.

Derivation in Czech is characterized by homonymy (polyfunctionality)[2] of affixes and, at the same time, by their synonymy. Many affixes convey more than one meaning, cf. the suffix *-ka* deriving the diminutive noun *vlnka* 'small wave' from *vlna* 'wave', the female noun *hráčka* 'female player' derived from *hráč* 'player', the agent noun *mluvka* 'talker' from *mluvit* 'to talk', or the location noun *skládka* 'dump' from *skládat* 'to dump'. From the opposite perspective, a particular meaning is usually expressed by several formally different affixes, cf. the suffixes *-ka* in *stavitelka* 'female builder' derived from *stavitel* 'builder', *-yně* in *kolegyně* 'female colleague' from *kolega* 'colleague', *-ice* in *lékarnice* 'female pharmacist' from *lékárník* 'pharmacist', and *-ová* in *švagrová* 'sister-in-law' from *švagr* 'brother-in-law' for female nouns.

The size of the DeriNet data as well as the fact that the database is still under construction were the main reasons why semantic labels were not assigned manually but a Machine Learning experiment was designed for this task. Five semantic labels were included into this pilot experiment, namely DIMINUTIVE, POSSESSIVE, FEMALE, ITERATIVE, and ASPECT. While the former four labels correspond to semantic concepts proposed for comparative research into affixation (Bagasheva, 2017), the latter label (ASPECT) was introduced to apply to suffixation of verbs that does not affect the lexical meaning but changes the category of aspect (from imperfective to perfective, or the other way round).[3]

Training and test data for the Machine Learning experiment, containing both positive and negative examples of the five labels to assign, were compiled by exploiting several language resources and reference grammars of Czech (cf. Ševčíková and Kyjánek in press for details).

Using morphological categories and character n-grams of both the base words and the derivatives as features and multinomial logistic regression as method, precision and recall achieved in the Machine Learning task (each above 96 %) indicate that the derivational families organized into rooted trees and the features included provide a sufficient basis for resolving the homonymy and synonymy of affixes in most cases. An analysis of incorrectly labelled relations pointed out, for example, to feminines incorrectly assigned the FEMALE label such as *profesura* 'professorship' (derived from *profesor* 'professor') and *krejčovna* 'tailor's workshop' (from *krejčí* 'tailor'); these particular problem could be solved by introducing the animacy feature to feminine nouns because the label is intended to be assigned only with female counterparts of masculines. The resulting annotation of approx. 150 thousand labels was included into the DeriNet 2.0 data. See Table 3 for a breakdown of the counts of the different categories.

2.4 Compounds

In the previous versions of DeriNet, compounding could not be adequately modelled due to the highly constrained data structure used as it allowed to specify a single base word for each derivative. In DeriNet 2.0, we introduce the notion of multi-node relations, which allow specifying any number of parent and child lexemes. Compounding is then annotated as a relation with multiple parent lexemes. For technical reasons, a single parent and a single child must always be marked as the main ones. For example, the adjective *jihoruský* 'south-Russian' points to the adjective *ruský* 'Russian' by the main-parent link

[2]The terms "homonymy" / "polyfunctionality" are preferred to "polysemy" in the recent accounts (Karlík et al., 2012; Šimandl, 2016).

[3]As formation of aspectual pairs exploits derivational affixes in Czech, the decision has been made to model this process as deverbal derivation in the DeriNet database (Ševčíková et al., 2017).

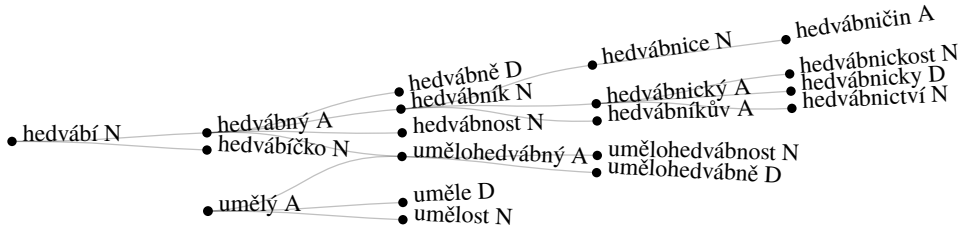

Figure 2: The derivational family of the lexeme *hedvábí* ‘silk’ and a tiny excerpt from the family of the lexeme *umělý* ‘artificial’.

and to the noun *jih* ‘south’ by a non-main-parent link. The interfix *-o-* is often added between the bases in compounds in Czech.

DeriNet 2.0 contains only a small sample of such compound annotations, serving, again, rather as a proof of concept. Out of around 33 thousand lexemes that were labelled as compounds in DeriNet 1.7 (just by a value of a binary flag, without their compositional parents being identified), we extracted 723 lexemes whose parents can be guessed automatically with relatively high reliability using just a set of string-based heuristics. Subsequently we checked the list manually, which resulted in 600 compounds for which both compositional parents are captured in DeriNet 2.0.

The procedure for guessing the parents works as follows: First, decompose the lemma of a known compound by finding an ‘o’ in it and extracting the substrings preceding and following it. The first substring is looked up in the dictionary as-is or amended by appending ‘ý’, ‘í’, ‘y’, ‘i’, ‘o’ or ‘a’ (these are common inflectional suffixes and word-final characters in Czech). The second substring is looked up in the dictionary verbatim. If these lookups result in finding only a single pair of candidate parent lemmas, output them, otherwise (if there are no matches or several) end the procedure without producing any output. This selection process is highly biased, as it selects only lexemes whose parents can be conclusively detected by simple string manipulation and ignores ambiguous cases.

2.5 Fictitious lexemes

When climbing from a derived word up to its base parent and continuing upwards, we should ideally end up in a tree root whose lemma is unmotivated (in the synchronous sense, i.e. there is no parent in the contemporary language). However, in some cases there is a strong intuition that a virtual node (corpus- or dictionary unattested) would be helpful, as it would complete a certain analogy pattern. For instance, one is tempted to add a non-existent lemma *bízet*, as it would naturally serve as a derivational base for *nabízet* ‘to offer’, *vybízet* ‘to prompt’, *pobízet* ‘to urge’ and others. In other configurations, a virtual lemma such as *tmívat* could serve as an intermediate node connecting a (corpus-attested) lemma *stmívat se* ‘to get dark’ with its (corpus-attested) grand-parent *tma* ‘darkness’, as the derivation is (again, by analogy to other derivational clusters) perceived as two-phase. We call such artificially added lexemes *fictitious lexemes*. As a proof of concept, we added 13 such lexemes into DeriNet 2.0, which allowed adding 41 derivations for prefixed verbs that should clearly not remain in tree root positions.

Our approach to fictitious lexemes is related to the linguistic discussion on cranberry morphemes (Aronoff, 1976) and, more recently, on paradigm gaps (e.g. Stump 2019). However, the basic building unit of DeriNet is still a lexeme, not a morpheme, and thus there is no technical means e.g. for expressing that a set of prefixed verbs makes use of the same morpheme.

3 New data format

Previous versions of DeriNet were published in a simple tab-separated-values text database file, which contained a lemma, part of speech and an optional link to the derivational parent on each line; see Table 4 for an excerpt from DeriNet 1.7. None of the new features can be represented in the old format, and so a new one was required. The old format cannot be easily extended in a backwards-compatible way, as there is no reserved field that identifies the version and the only possible simple extension – adding new columns to the end of each line – is not compatible with existing tooling that uses several extra columns

ID	Lemma	Dictionary ID	POS	Parent ID
205205	hedvábíčko	hedvábíčko	N	205206
205206	hedvábí	hedvábí	N	
205207	hedvábně	hedvábně_(*1ý)	D	205219
205208	hedvábnice	hedvábnice_(*3ík)	N	205215
205209	hedvábničin	hedvábničin_(*3ce)	A	205208
205211	hedvábnickost	hedvábnickost_(*3ý)	N	205213
205212	hedvábnicky	hedvábnicky_(*1ý)	D	205213
205213	hedvábnický	hedvábnický	A	205215
205214	hedvábnictví	hedvábnictví	N	205213
205215	hedvábník	hedvábník	N	205219
205216	hedvábníkův	hedvábníkův_(*2)	A	205215
205218	hedvábnost	hedvábnost_(*3ý)	N	205219
205219	hedvábný	hedvábný	A	205206
...	...	...	...	...
768083	umělohedvábně	umělohedvábně_^(*1ý)	D	768085
768084	umělohedvábnost	umělohedvábnost_^(*3ý)	N	768085
768085	umělohedvábný	umělohedvábný	AC	
...	...	...	...	...
768106	umělý	umělý	A	768197
768020	uměle	uměle_(*1ý)	D	768106

Table 4: The tree below the word "hedvábí" (silk) and excerpts of two related trees in DeriNet 1.7. Since compounding cannot be annotated in this format, the word *umělohedvábný* 'made of artificial silk' is marked as a compound using the 'C' mark in the part-of-speech category (fourth) column, but it is not connected to its parents *umělý* 'artificial' and *hedvábný* 'made of silk'. The Dictionary ID column lists the lemma together with technical suffixes as used by the MorfFlex dictionary – these are stored in DeriNet to allow interlinking the two resources.

for debugging information. Therefore, as compatibility with existing tools has to be broken anyway, we decided to create the new format from the ground up. When designing it, we drew inspiration from the CoNLL-U format (Nivre et al., 2016), which recently became a widely used representation of syntactic annotation.

The new format is still textual and lexeme-based, but it allows for a wider range of annotations. In addition to the lemma and part-of-speech tag, each lexeme can be annotated by key-value pairs specifying its properties (e.g. the morphological categories), a list of its morphemes together with their properties, and by any number of directed word-formation relations. Each relation can connect multiple parents with multiple children, and so the format can express one-to-one relation such as derivation or conversion, as well as many-to-one relations such as compounding. The relations are stored together with their children, connecting them to their parents, but otherwise behave like separate entities, and they can also be annotated with arbitrary key-value pairs (e.g. the semantic labels). Furthermore, there is space for custom (possibly language-specific) extensions of the format in the form of JSON-encoded data (Bray, 2017) stored in the last column. See Table 5 for an excerpt from DeriNet 2.0 showing the new format and Figure 2 for a visualization of this data.

The key-value pairs are serialized into textual form by joining each pair by an equals sign and concatenating all such pairs describing a single entity with ampersands: `key1=value1&key2=value2`. If the field in question describes multiple entities, such as the segmentation, the different entities are concatenated with vertical bars: `key1=value1&key2=value2|keyA=valueA`.

To simplify processing of the data, which has the form of a general graph, we explicitly select tree-shaped substructures from the graph and store the corresponding "main parent" IDs in a dedicated column. The lexemes in the file are grouped according to these trees, which correspond to derivational families, with compounds added to the family of one of its parents. This enables e.g. performing a depth-first search over the structure of the derivational families without having to explicitly avoid cycles by marking

ID	Language-specific ID	Lemma	POS	Morphological features	Morpheme segmentation	Main parent ID	Parent relation
144293.0	hedvábí#NNN??-----A---?	hedvábí	N	Gender=Neut			
144293.1	hedvábný#AA???----??---?	hedvábný	A			144293.0	Type=Derivation
144293.2	hedvábně#Dg-------??---?	hedvábně	D			144293.1	Type=Derivation
144293.3	hedvábník#NNM??-----A---?	hedvábník	N	Animacy=Anim &Gender=Masc		144293.1	Type=Derivation
144293.4	hedvábnice#NNF??-----A---?	hedvábnice	N	Gender=Fem		144293.3	SemanticLabel=Female &Type=Derivation
144293.5	hedvábničin#AU????--------?	hedvábničin	A	Poss=Yes		144293.4	SemanticLabel=Possessive &Type=Derivation
144293.6	hedvábnický#AA???----??---?	hedvábnický	A			144293.3	Type=Derivation
144293.7	hedvábnickost#NNF??-----?---?	hedvábnickost	N	Gender=Fem		144293.6	Type=Derivation
144293.8	hedvábnicky#Dg-------??---?	hedvábnicky	D			144293.6	Type=Derivation
144293.9	hedvábnictví#NNN??-----A---?	hedvábnictví	N	Gender=Neut		144293.6	Type=Derivation
144293.10	hedvábníkův#AU???M--------?	hedvábníkův	A	Poss=Yes		144293.3	SemanticLabel=Possessive &Type=Derivation
144293.11	hedvábnost#NNF??-----?---?	hedvábnost	N	Gender=Fem		144293.1	Type=Derivation
144293.12	umělohedvábný#AA???----??---?	umělohedvábný	A			144293.1	Sources=195833.258,144293.1 &Type=Compounding
144293.13	umělohedvábnost#NNF??-----?---?	umělohedvábnost	N	Gender=Fem		144293.12	Type=Derivation
144293.14	umělohedvábně#Dg-------??---?	umělohedvábně	D			144293.12	Type=Derivation
144293.15	hedvábíčko#NNN??-----A---?	hedvábíčko	N	Gender=Neut		144293.0	SemanticLabel=Diminutive &Type=Derivation
...	...	...	...	...	...	...	...
195833.258	umělý#AA???----??---?	umělý	A		End=2 &Morph=um &Start=0 &Type=Root	195833.4	Type=Derivation
195833.259	uměle#Dg-------??---?	uměle	D		End=2 &Morph=um &Start=0 &Type=Root	195833.258	Type=Derivation

Table 5: The lexeme *hedvábí* 'silk' and derivationally related lexemes (i.e. a derivational family represented as a tree) in DeriNet 2.0. The last column containing language- and resource-specific data has been omitted; in Czech DeriNet 2.0, it contains the technical dictionary ID for linking with MorfFlex and the "compound yes/no" flag from previous versions of DeriNet. The line with dots divides the derivational family of *hedvábí* 'silk' from that of *umělý* 'artificial', which is the second base word for the compound *umělohedvábný* 'made of artificial silk'.
The family containing the lexeme *umělý* is large enough to have been included in the annotation of root morphemes. This annotation is present in the sixth column. The family of *hedvábí* is not annotated yet and its sixth column is therefore empty.

visited lexemes, as it guarantees that a search starting from the base lexeme of the family will visit every lexeme in it exactly once. There are no restrictions on the relations not participating in the tree-shaped substructure, so it is possible to annotate double motivation and other general word-formation structures.

Inside the database, all lexemes are unambiguously specified using an ID. The IDs are hierarchical: they are composed of the number of the tree they are in, followed by the number of the lexeme in the tree. These IDs are used to specify the endpoints of relations. Because the hierarchical numerical IDs are opaque and they change when a lexeme is reconnected, a more permanent identification of a lexeme is possible using a field reserved for this purpose. In the Czech data, this fields contains the lemma and the tag mask introduced above.

Detailed documentation of the file format and the tools created to process it is available in the `doc/` directory of the DeriNet repository at https://github.com/vidraj/derinet.

4 Conclusions

The DeriNet database was enriched with several different kinds of information about the lexemes and

relations contained therein, which were previously missing. The newly added annotation is useful or even required for many tasks, e.g. the availability of morphological categories was vital to annotating the relations with semantic labels, and the annotation of root morphemes allowed us to cross-check the already present derivational relations with another source of information.

The format we developed for storing and distributing the resulting network is supposed to be general, extensible and language-agnostic enough to be usable by other projects as well. By using a common format, the different networks can benefit from a shared set of tools and services and their users can more easily compare their properties, and through that hopefully also the properties of different languages.

Acknowledgments

This work was supported by the Grant No. GA19-14534S of the Czech Science Foundation, by the Charles University Grant Agency (project No. 1176219) and by the SVV project number 260 453. It has been using language resources developed, stored, and distributed by the LINDAT/CLARIAH CZ project (LM2015071, LM2018101).

References

Mark Aronoff. 1976. *Word Formation in Generative Grammar*, volume 1 of *Linguistic inquiry monographs*. MIT Press, Cambridge, Massachusetts, USA.

R. Harald Baayen, Richard Piepenbrock, and Leon Gulikers. 1995. CELEX2. Linguistic Data Consortium, Catalogue No. LDC96L14. https://catalog.ldc.upenn.edu/LDC96L14.

Alexandra Bagasheva. 2017. Comparative Semantic Concepts in Affixation. In *Competing Patterns in English Affixation*. Peter Lang, Bern, Switzerland, pages 33–65.

Tim Bray. 2017. The JavaScript Object Notation (JSON) Data Interchange Format. RFC 8259.

Jan Hajič and Jaroslava Hlaváčová. 2013. MorfFlex CZ. http://hdl.handle.net/11858/00-097C-0000-0015-A780-9.

Nabil Hathout and Fiammetta Namer. 2014. Démonette, A French Derivational Morpho-Semantic Network. *Linguistic Issues in Language Technology* 11:125–162.

Petr Karlík et al. 2012. *Příruční mluvnice češtiny*. NLN, Prague, Czech Republic.

Mikko Kurimo, Sami Virpioja, Ville T. Turunen, Graeme W. Blackwood, and William Byrne. 2009. Overview and results of Morpho Challenge 2009. In *Workshop of the Cross-Language Evaluation Forum for European Languages*. Springer, pages 578–597.

Eleonora Maria Gabriella Litta Modignani Picozzi, Marco Carlo Passarotti, and Chris Culy. 2016. Formatio Formosa est. Building a Word Formation Lexicon for Latin. In *Proceedings of the 3rd Italian Conference on Computational Linguistics*. pages 185–189.

Joakim Nivre et al. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*. ELRA, pages 1659–1666.

Gregory Stump. 2019. Some sources of apparent gaps in derivational paradigms. *Morphology* 29:271–292.

Jonáš Vidra, Zdeněk Žabokrtský, Lukáš Kyjánek, Magda Ševčíková, and Šárka Dohnalová. 2019. DeriNet 2.0. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. http://hdl.handle.net/11234/1-2995.

Magda Ševčíková, Adéla Kalužová, and Zdeněk Žabokrtský. 2017. Identification of Aspectual Pairs of Verbs Derived by Suffixation in the Lexical Database DeriNet. In *Proceedings of the Workshop on Resources and Tools for Derivational Morphology*. EDUCatt, Milan, Italy, pages 105–116.

Magda Ševčíková and Lukáš Kyjánek. in press. Introducing Semantic Labels into the DeriNet Network. *Jazykovedný časopis* .

Magda Ševčíková and Zdeněk Žabokrtský. 2014. Word-Formation Network for Czech. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*. ELRA, Reykjavik, Iceland, pages 1087–1093.

Josef Šimandl, editor. 2016. *Slovník afixů užívaných v češtině*. Karolinum, Prague, Czech Republic.