PBML

The Prague Bulletin of Mathematical Linguistics

NUMBER 118 APRIL 2022

## EDITORIAL BOARD

Published twice a year by Charles University (Prague, Czech Republic)

Editorial office and subscription inquiries:
ÚFAL MFF UK, Malostranské náměstí 25, 118 00, Prague 1, Czech Republic
E-mail: `pbml@ufal.mff.cuni.cz`

ISSN 0032-6585

PBML

The Prague Bulletin of Mathematical Linguistics

NUMBER 118 APRIL 2022

# CONTENTS

## Articles

The Prague Bulletin of Mathematical Linguistics
NUMBER 118 APRIL 2022 5–23

# Interpreting Statistical Models for Denominal Adjective Formation in Russian

Natalia Bobkova

CLLE, Université de Toulouse 2 Jean Jaurès, France

**Abstract**

This study focuses on cases of suffixal rivalry in denominal adjective formations in Russian, namely on two adjectival suffixes: *-n-* and *-sk-*. We use statistical modelling (multivariate logistic regression) to shed light on properties of base nouns that contribute to the choice of one of the competing suffixes. In the first part, we provide model interpretation through traditional metrics (accuracy, confusion matrix and model coefficients with their respective *p*-values). However, model accuracy may not be uniform if we compare different samples of the data set and may take a wide range of values. In the second part of this study, we complete our interpretation of model results by performing error analysis in order to get a better understanding of the underlying properties of base nouns that cause model failure. We explore Responsible AI Toolbox widgets for this purpose. One main result of this study is that the same semantic base noun properties are related to both high model performances and model errors.

## 1. Introduction

The derivation of adjectives from nouns is a complex process in Russian morphology, as there is a lot of variation in the range of suffixes employed. Hence, they constitute a good testing ground for the study of the competition between rival derivational strategies for the same syntactic and semantic function (Lindsay and Aronoff, 2013; Aronoff, 2016).

The use of quantitative methods to investigate the situations of affix rivalry has increased recently. The studies rely heavily on statistical and computational methods as opposed to traditional qualitative research. Quantitative methods are exploited to evaluate the influence of different factors on the selection of rival affixes. Inferential statistics can be based on a variety of models (Baayen et al., 2013), including ana-

Corresponding author: natalia.bobkova@univ-tlse2.fr

Cite as: Natalia Bobkova. Interpreting Statistical Models for Denominal Adjective Formation in Russian. The Prague Bulletin of Mathematical Linguistics No. 118, 2022, pp. 5–23. doi: 10.14712/00326585.017.

logical models (Chapman and Skousen, 2005; Arndt-Lappe, 2014), logistic regression (Bonami and Thuilier, 2018), word vectors (Wauquier, 2020; Guzmán Naranjo and Bonami, 2021; Huyghe and Wauquier, 2021), neural networks (Guzmán Naranjo, 2019; King et al., 2020).

The competition between adjectival suffixes is determined by a complex combination of phonological, morphological and semantic factors. In this paper we aim at modeling suffixal rivalry in the construction of denominal adjectives in Russian. The approach adopted in this paper consists in studying non-ambiguous cases for each suffix in the data set and in highlighting the emerging properties of base nouns that allow to tease apart competing suffixes. For illustration purposes we will use *-n-* and *-sk-* suffixes data, however, the approach can be applied for both binary and multiclass classification problems (i.e. to include more than two suffixes in the study).

The goal is to understand the role of base noun properties in predicting *-n-* and *-sk-*. There is a variety of models which can be used for this purpose due to their high interpretability. For instance, logistic regression, decision trees or random forest can output variable importance scores (base noun properties) in explaining the outcome (suffix). In this study we use multivariate logistic regression, a well-established statistical modelling framework. The choice of logistic regression over other models is driven by several factors: it is a tool based on statistical formulae, the direction of coefficients (positive or negative) can be associated with two classes of binary classification and, finally, the coefficients are accompanied by statistical significance tests (with p-values). Even if this model has all the advantages listed above, we will not limit our investigation to the classical tools in order to understand it (such as its table of coefficients). In this paper we will explore various quantitative methods for error analysis aiming to highlight patterns or combination of patterns which are not captured by our model, and the reasons behind them.

The error analysis was proposed by King et al. (2020) as approach to understand the output of sequence-to-sequence models, which are generally hard to interpret, for inflectional task in Russian. This paper goes further and uses quantitative and qualitative approaches for error analysis and model interpretation. Based on error analysis, our study provides a new perspective on the nature of suffix rivalry in Russian derivation and sheds light on previously unseen phenomena.

The data on which our study is performed were extracted from the Russian National Corpus. The data set is composed of highly frequent adjectives. Section 2 discusses different problems which emerge when studying adjectives in Russian. Section 3 presents the overview of the Russian National Corpus, data set constitution and base noun properties annotation. Section 4 focuses on building a logistic regression classifier, it provides data on its performance as well as model summary highlighting the base noun properties which are statistically significant for classification task. Section 5 focuses on error analysis and diagnostics, sheds light on base noun properties which may be misleading for the model and discusses the underlying reasons for errors. The error analysis here is complementary to the logistic regression task.

## 2. Adjectives in Russian

There are various strategies to derive adjectives from nouns in Russian. Classical grammars such as Townsend (1975) or Švedova (1980), for instance, enumerate more than 25 suffixes, which have different degrees of productivity. Three suffixes are identified as being productive in synchrony (Zemskaya, 2015; Hénault and Sakhno, 2015; Kustova, 2018): *-n-*, *-sk-* and *-Ov-* (capital *O* in both cases represents a vowel that may correspond, phonologically, to different surface forms, and orthographically to <o> or <e>). The suffixes in question can be considered as the three main adjectival suffixes (abstract entities, denoted in capital letters), while others may be interpreted as their extended variants, denoted in small letters (Bobkova and Montermini, 2019):

- -N-: *-n-*, *-Ovn-*, *-ičn-*, *-ivn-*, *-on(n)-*, *-en(n)-*, *-(e)stven(n)-*, *-ozn-*, *-al'n-*, *-onal'n-*, *-arn-*, *-in-*;
- -SK-: *-sk-*, *-esk-*, *-česk-*, *-ičesk-*, *-ističesk-*, *-ijsk-*, *-ansk-*, *-ensk-*, *-insk-*, *-istsk-*, *-Ovsk-*;
- -OV-: *-Ov-*.

Recent developments in derivational morphology, cf. Hathout (2011); Plénat (2011); Roché (2011) among others, consider that various types of constraints (phonological, morphological, semantic, pragmatic, etc.) display a complex interaction, resulting in the choice of one of the rival suffixes. However, in the existing literature on Russian language the choice of one or the other suffix is often studied theoretically, through extended data, but not necessarily by means of quantitative analysis. For instance, in Townsend (1975); Švedova (1980); Zemskaya (2015) we can encounter extensive indications on phonological, semantic or lexico-morphological factors that allow the combination with each suffix in question. Graščenkov (2019) references Švedova for the properties of base nouns discussed above and studies syntactic properties of suffixes *-n-* and *-sk-*.[1] Graudina et al. (2001); Hénault and Sakhno (2015), for instance, focus on the semantics of derived adjectives and provide evidence on distinction between *-n-* and *-sk-* based on context the adjectives appear in. However, all the indications are not supported with quantitative and/or statistical evidence.

For the purposes of the present study we will focus on phonological, morphological and semantic properties of the base nouns.

The examples of nouns combining with *-n-* are given in Table 1.[2]. In Švedova (1980), for instance, the following non-extensive indications on *-n-* can be encountered. Semantically, this suffix mainly combines with non-animate common nouns, either abstract (1) or concrete (2), although animate nouns are also possible bases (3). Phonologically, it is stress-neutral, as it combines both with bases with stress on

[1] The analysis is based on the ability of *-n-* and *-sk-* adjectives to form adverbs, to have short forms and comparative forms in their paradigms, to derive abstract nouns, to combine with evaluative suffixes.

[2] For illustration purposes we provide stress position information for the base nouns in Tables 1 and 2 In the rest of the paper these indications will be excluded, except if relevant.

the stem (4) or on inflection (5), and it selects stems displaying consonant mutation (6, 7). Etymologically, it combines both with native (8) and foreign (9) bases.

| | **noun** | **adjective** | **gloss** |
|---|---|---|---|
| 1 | *gnev* | *gnevn*(*yj*) | 'anger' |
| 2 | *kiparís* | *kiparisn*(*yj*) | 'cypress' |
| 3 | *inženér* | *inženern*(*yj*) | 'engineer' |
| 4 | *kómnat*(*a*) | *kómnatn*(*yj*) | 'room' |
| 5 | *zim*(*á*) | *zímn*(*ij*) | 'winter' |
| 6 | *jazýk* | *jazyčn*(*yj*) | 'tongue / language' |
| 7 | *drug* | *družn*(*yj*) | 'friend' |
| 8 | *dym* | *dymn*(*yj*) | 'smoke' |
| 9 | *arxitektúr*(*a*) | *arxitekturn*(*yj*) | 'architecture' |

*Table 1. Sample with -n- suffixation*

Table 2 provides examples of nouns combining with *-sk-*. This suffix does not seem to be selective semantically, since it may combine with inanimate (1) and animate (2) nouns, including nouns denoting humans (3), and may also combine with proper nouns (4). Phonologically, it privileges stems ending in alveolar (5) or dental (6) consonants, and, like *-n-*, it selects nouns with stress on the stem, and mutated stems (7,8).

| | **noun** | **adjective** | **gloss** |
|---|---|---|---|
| 1 | *universitét* | *universitetsk*(*ij*) | 'university' |
| 2 | *kon'* | *konsk*(*ij*) | 'horse' |
| 3 | *bandít* | *banditsk*(*ij*) | 'bandit' |
| 4 | *Irán* | *iransk*(*ij*) | 'Iran' |
| 5 | *soséd* | *sosedsk*(*ij*) | 'neighbour' |
| 6 | *šef* | *šefsk*(*ij*) | 'boss' |
| 7 | *Vólg*(*a*) | *volžsk*(*ij*) | 'Volga (river)' |
| 8 | *Čéxi*(*ja*) | *češsk*(*ij*) | 'Czechia' |

*Table 2. Sample with -sk- suffixation*

The literature revision proves that the indications on these properties often lack precision: the same base noun property can be listed as favourable for different suffixes. It remains unclear which properties are statistically significant for the suffix choice. The goal of this study is twofold: first, we will provide statistic evidence on the base noun properties that allow to discriminate between *-n-* and *-sk-* for highly frequent adjectives through logistic regression model. Second, we will identify and diagnose in depth the error patterns; this investigation will shed light on the distribution of base noun properties across different subsets of data which are prone to model failure.

## 3. Data

To perform our analysis, we proceeded with web scraping adjectives from the Russian National Corpus (Plungjan et al., 2005),[3] a corpus of modern Russian containing over 600 million words. This corpus is divided in several subcorpora. For the purpose of this study we are interested in standard Russian, both written and spoken. Consequently, the adjectives were extracted from five subcorpora: main (texts representing standard Russian: modern written texts from the 1950s to the present day, real-life Russian speech recordings from the same period, and early texts from the middle of the 18th to the middle of the 20th centuries), media (articles from mass media between 1990 and the 2000s), multimedia (Russian movies between 1930 and 2000), spoken (recordings of public and spontaneous spoken Russian and the transcripts of the Russian movies) and poetic (covers the time frame between 1750 and the 1890s, but also includes some poets of the 20th century).

Having established the types of subcorpora we are interested in, we automatically extracted adjectives by searching lemmas with a final sequence corresponding to *-n-* or *-sk-* immediately preceding inflectional suffixes typical of citation forms of Russian adjectives.[4] 78113 lemmas were extracted, we automatically filtered extended variants (almost 1/3 of the data set). Semi-automatic and manual cleaning further allowed to discard >70% false positives, e.g. forms corresponding to adverbs derived with *-n-* (*vnezapno* 'suddenly'), possessive adjectives in *-in* (*mamin* 'mother$_{\text{POS}}$'), proper nouns (surnames) ending in *-sk-* (*Stanislavsk*(*ij*) 'Stanislavsky'). This first list was additionally filtered in order to keep only adjectives clearly derived from nouns. The vast majority of remaining adjectives are denominal, other cases were removed: noun to adjective conversions (*zdorov'*(*e*) - *zdorov*(*yj*) 'health'; *tajn*(*a*) - *tajn*(*yj*) 'secret'), adverb to adjective conversions (*dëševo* - *dešëv*(*yj*) 'cheap', *rano* - *rann*(*ij*) 'early'), as well as the adjectives without any motivating base. Furthermore, we only took into account adjectives having token frequency >100, excluding non frequent formations along with hapaxes from the present study.

[3] Available at https://ruscorpora.ru/. The choice of web scraping method is driven by the absence of an official API for data access in Ruscorpora.

[4] The citation form of adjectives corresponds to nominative masculine singular. Three orthographic forms are possible: <yj>, <ij>, <oj>.

Base nouns were also automatically reconstructed for each adjective. In case of multiple base candidates (*zritel'/zreni(e)* - *zritel'n(yj)* 'viewer/vision') and polysemy (*kamer(a)*$_1$/*kamer(a)*$_2$ - *kamern(yj)* 'cell/chamber'), these potential base nouns, as well as nouns with different semantics, were included as separate entries and annotated accordingly. Manual assessment at this stage led to verification of the exact shape of the reconstructed base nouns. The final data set was composed of 1048 types (620 for *-n-* and 428 for *-sk-*).

The competition between affixes is driven by a complex combination of factors. In order to examine different dimensions of rivalry, we annotated several properties of base nouns that have been highlighted in previous linguistic works as potential predictors of the suffix, as discussed in Section 2. In what follows we will present these properties in details and give a brief overview of the studies of rivalry mainly in English and French that use the same properties as predictors in modelling.

Etymological property include one binary predictor:

- `Source`: whether the base noun is of Slavic (`0`) or foreign (`1`) origin.

Phonological properties include information about the following features:

- `LastP`: the last phoneme of the stem (`Lab`: labial, `Den`: dental, `Alv`: alveolar, `Vel`: velar or `Vow`: vowel);
- `SyllB`: the length of the base noun in syllables - the only continuous property in the dataset;
- Stress position is also taken into consideration:
  - `AccSyl`: from the phonological point of view: which syllable is stressed – `D`: ultimate, `Ad`: penultimate, `Aad`: antepenultimate (*zim(á)* 'winter', *víšn(ja)* 'cherry', *rádug(a)* 'rainbow');
  - `AccPos`: from the morphological point of view: if the stress is positioned on `R`: the root of the base noun, or – if any – `S`: derivational or `F`: inflectional suffix (*son* 'dream', *marksízm* 'marxism', *galav(á)* 'head').

Both the last phoneme of the stem and the length of base noun in syllables are highlighted as important in prediction of the suffix by Lignon (2010) and Bonami and Thuilier (2018) in French, by Lindsay and Aronoff (2013) in English. We complete the list of phonological properties with information on stress position since it is not fixed in Russian and may influence the choice of the suffix.

Morphological properties include only one predictor :

- `InflCl`: the inflectional class of base nouns which is represented by the I, II or III inflectional class (*pap(a)*$_{\text{I.M}}$ 'dad', *pesn(ja)*$_{\text{I.F}}$ 'song'; *stol*$_{\text{II.M}}$ 'table', *del(o)*$_{\text{II.N}}$ 'business'; *ten'*$_{\text{III.F}}$ 'shadow').

We follow a canonical distinction between 3 inflectional classes, although Russian nouns may be divided into larger sets of classes and subclasses (Zaliznjak, 2003; Parker and Sims, 2019; Guzmán Naranjo, 2020). We only include inflectional class as morphological property, however, morphological structure of base nouns may be interesting as well to study suffix rivalry further (Missud and Villoing, 2020; Varvara, 2020).

Morpho-phonological allomorphies typical of Russian inflection and derivation were annotated as well. They include such properties as:

- *Vowel0*: vowel / Ø alternation, binary property (*dvorec* - *dvorcov*(*yj*) 'palace');
- *ConsM*: consonant mutation, binary property (*tvorog* - *tvorožn*(*yj*) 'cottage cheese').

Both vowel alternation and consonant mutation reflect diachronic processes in Russian and do not correspond to synchronically productive phonological phenomena (Kapatsinski, 2010; Sims, 2017; Timberlake, 2004).

Possible differences in the semantics of derivatives may be considered as well, with respect to descriptive properties (Baeskow, 2012; Fradin, 2016). We include the following semantic properties of base nouns in this study:

- Binary distinct properties of [±proper], [±human], [±animate], [±concrete], [±countable];
- `A`: animacy, or the combination of the properties listed above into five groups (Thuilier, 2012):
  - `PropHum`: proper human (*Pifagor* 'Pithagoras');
  - `ComHum`: common human/animate (*sobak*(*a*) 'dog');
  - `ComConc`: common concrete (*dom* 'house');
  - `PropNHum`: proper non-human (*Al'p*(*y*) 'Alps');
  - `ComAbst`: common abstract (*sojuz* 'alliance').

After performing descriptive statistics analysis and test for multicolinearity,[5] some data were removed before modeling. For instance, the nouns with samples of properties that are not large enough to be statistically representative were dropped out (nouns with six-syllabic structure, nouns where the forth syllable from the end is stressed). Highly correlated base noun properties were also removed. This concerns binary semantic features since they strongly correlate to animacy subclasses, as well as consonant mutation which strongly correlates to velar ending stems.

The data set for modelling is composed of 1020 examples, 612 for -*n*- and 408 for -*sk*-.

## 4. Model

All the base noun properties listed in previous section virtually combine to form a complete picture of situations of rivalry. In what follows we will examine their predictive power for the suffix choice when they are put all together.

We use logistic regression, a multifactorial statistical tool which allows to examine the relationship between a binary dependent categorical variables and predictor

[5] For more details on methodological aspects cf. Bobkova (2022).

variables. The implementation is made with statsmodels module in Python (Seabold and Perktold, 2010).

The data were randomly divided into training and test (with test size of 20%, so the model was trained on 816 examples and tested on 204). We ran 500 simulations of train-test split with a different random state.[6] The goal of this manipulation is twofold. First, we aimed at assessing overall model AUC score when trained and tested of different subsets of original data (`mean AUC`: 0.8957, `min AUC`: 0.8345; `max AUC`: 0.9502, `std`: 0.020)).[7] Second, since overall model performance is high it does not make a lot of mistakes, and given the relatively small test set, we searched for the worst performing model in order to maximize error rates and have enough material for further analysis.

We will now focus on the model with the lowest AUC (0.8345) and investigate its performance and properties. We will use logistic regression table of coefficients (Table 3) to evaluate statistical significance of predictors.

First, we use p-values in order to understand if a particular base noun property is useful for suffix prediction. The p-value less than 0.05 suggests that the property has a significant effect on the suffix choice. The model summary states that [+common, +human] and [-common, +human] semantic properties, as well as [+dental]-ending stems are statistically significant for predicting suffix (p<0.000). The following parameters are also significant, but to a lesser extent: [+labial]- (p<0.012), [+alveolar]- (p<0.032) and [+velar]-ending stems (p<0.042), inflectional class 2 (p<0.021) and 1 (p<0.031). Source, the length of base noun in syllables, vowel-∅ alternation, [+common, +concrete] semantic property, morphological and phonological stress positions are not statistically significant for *-n-* and *-sk-* classification problem.

Second, we can interpret coefficients which compare the outcome for each level of a base noun property with the reference level (the reference levels for each categorical predictor correspond to Slavic origin, absence of Ø vowel, common abstract, stressed root, stressed antepenultimate syllable, inflectional class 3, vowel-ending stem). Positive coefficients increase the chances for the model to predict *-sk-* ([+common, +human], [-common, +human], inflectional class 1 and 2), negative coefficients, in turn, decrease odds for *-sk-* and increase the probability for predicting *-n-* ([+dental], [+labial]-, [+alveolar]- and [+velar]-ending stems).

Table 4 provides confusion matrix. 31 nouns out of 204 were misclassified, the error rate is 14.7%. This table also suggests that more classification errors were made for *-sk-* (25.3% of misclassified data) rather than for *-n-* (8.5% of errors). We will proceed with an in-depth investigation of these errors as well as underlying possible reasons for them in the following section.

[6] 500 is an arbitrary choice in order to have a large number of simulations.

[7] Compared to AUC score, overall accuracy score is higher: `mean accuracy`: 0.9079, `min accuracy`: 0.8534; `max accuracy`: 0.9559, `std`: 0.018.

| | coef | std err | z | P>\|z\| |
|---|---|---|---|---|
| **Intercept** | -4.0000 | 1.124 | -3.559 | 0.000 |
| **Source** | 0.0909 | 0.306 | 0.297 | 0.766 |
| **BaseLen** | 0.2158 | 0.184 | 1.173 | 0.241 |
| **Vowel0** | -1.0549 | 0.680 | -1.550 | 0.121 |
| **A_ComConc** | 0.2601 | 0.342 | 0.754 | 0.451 |
| **A_ComHum** | 4.2509 | 0.394 | 10.785 | 0.000 |
| **A_PropNHum** | 11.4359 | 1.177 | 6.461 | 0.000 |
| **StressMo_DerS** | 0.5390 | 0.584 | 0.924 | 0.356 |
| **StressMo_InfS** | -0.1964 | 0.732 | -0.268 | 0.789 |
| **StressPho_ad** | -0.7678 | 0.524 | -1.464 | 0.143 |
| **StressPho_d** | -0.4269 | 0.604 | -0.684 | 0.494 |
| **InflCl_1** | 2.7431 | 1.274 | 2.153 | 0.031 |
| **InflCl_2** | 2.8106 | 1.219 | 2.306 | 0.021 |
| **LastPh_cAlv** | -0.7735 | 0.361 | -2.143 | 0.032 |
| **LastPh_cDent** | -1.3741 | 0.379 | -3.627 | 0.000 |
| **LastPh_cLab** | -1.0472 | 0.416 | -2.518 | 0.012 |
| **LastPh_cVel** | -0.8042 | 0.396 | -2.031 | 0.042 |

Table 3. Model summary

| | predicted *-n-* | predicted *-sk-* |
|---|---|---|
| **true *-n-*** | 118 | 11 |
| **true *-sk-*** | 19 | 56 |

Table 4. Confusion matrix

Classification report is shown in Table 5. Even if the chosen model has the lowest accuracy, it still performs quite well: with accuracy of 85.3% and AUC of 83.5%, good precision and descent recall. However, these metrics, especially accuracy, may not be uniform across different subsets of data. Moreover, these metrics do not allow to identify important conditions of inaccuracies. The model may perform better for some initial base noun properties and worse for others. Therefore, an in-depth analysis is needed to convey a detailed interpretation of model behavior.

## 5. Error analysis

In this section we will further investigate the performance of the model, namely through data exploration and interpretability techniques as well as through an anal-

| metric | value |
|---|---|
| Accuracy | 0.853 |
| AUC | 0.835 |
| Precision | 0.836 |
| Recall | 0.747 |
| False Positive Rates | 0.085 |
| False Negative Rates | 0.253 |

*Table 5. Classification report*

ysis of how failure is distributed for a model. We will use visualisation methods provided by Responsible AI.[8]

The Error analysis[9] and Interpretability[10] dashboards are integrated within the Responsible AI Widgets. They enable a better understanding of overall and local predictions of the model as well as of model errors (Nushi et al., 2018; Amershi et al., 2019; Bansal et al., 2019; Srivastava et al., 2020). These tools allow to work with regression and classification problems, both binary and multiclass. Responsible AI tools can be used to assess any kind of models (statistical or machine learning), even the models which are not easily interpretable (for instance, deep learning models). In what follows we will complete the assessment of the logistic regression classifier used in this study.

Error analysis dashboard enables the visualization of data subsets with higher error rates than the overall error score. These errors may occur when the model faces specific set of properties among independent variables, i.e. the properties of base nouns for which the model underperforms.

As assessed in the previous section through confusion matrix, the overall error rate is 14.71% since 31 out of 204 base nouns were associated with the wrong suffix.

However error patterns may be complex and involve several properties of base nouns. The Figure 1 groups all misclassified data into subsets which can be easily interpreted in a tree-like structure. This tree uses the mutual information between each property and the error on the true labels to best separate error instances from success instances hierarchically in the data. This allows to visualize common patterns in model failure. The following information is available for this binary tree: error rate (portion of instances in the node for which the model is incorrect, shown through the

[8] https://github.com/microsoft/responsible-ai-toolbox

[9] https://github.com/microsoft/responsible-ai-toolbox/blob/main/docs/erroranalysis-dashboard-README.md

[10] https://github.com/microsoft/responsible-ai-toolbox/blob/main/docs/explanation-dashboard-README.md

intensity of color); error coverage (portion of all errors that fall into the node, shown through the fill rate of the node) and data representation (number of instances in the node, shown through the thickness of the incoming edge to the node along with the actual total number of instances in the node).

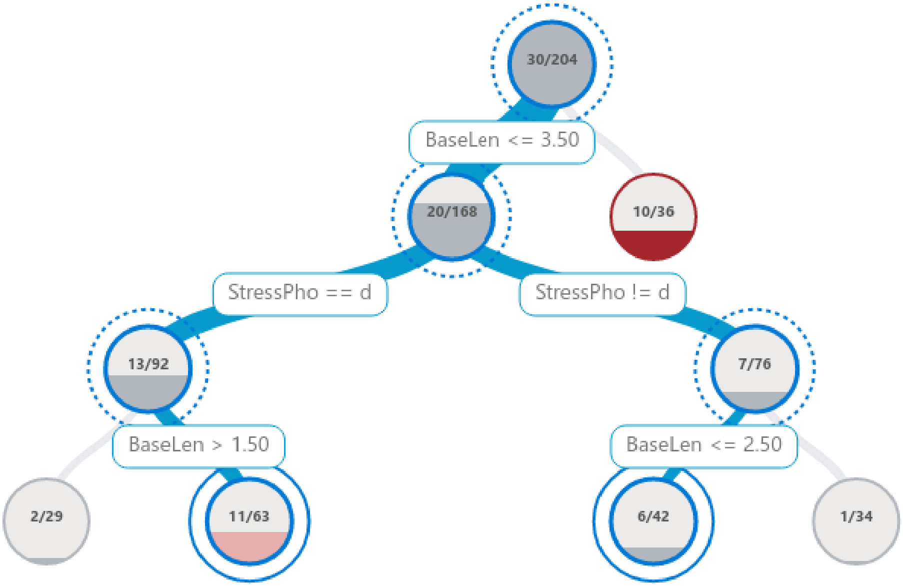

*Figure 1. Error tree for logistic regression model*

This decision tree represents combined data on two branches. The root node contains the information about the length of base noun in syllables. It allows for further partitioning data into two groups, based on the following condition: if the number of syllable is less than or greater than 3.5.

While the overall error rate is 14.71% for the whole dataset, the error rate can be as high as 27.78%, which corresponds to the extreme right branch with only one node, 10 out of 36 cases of wrong classification (for base nouns of 4 or 5 syllables). Six nouns are 5-syllabic (*gumanitarij* 'humanitarian', *bogoslovi*(*e*) 'theology', *artillerij*(*a*) 'infantry', *universitet* 'university', *žurnalistik*(*a*) 'journalism', *professional* 'professional'), the other four are 4-syllabic (*veterinar* 'vet', *vselennaj*(*a*) 'universe', *čudovišč*(*e*) 'monster', *distrib'jutor* 'distributor').

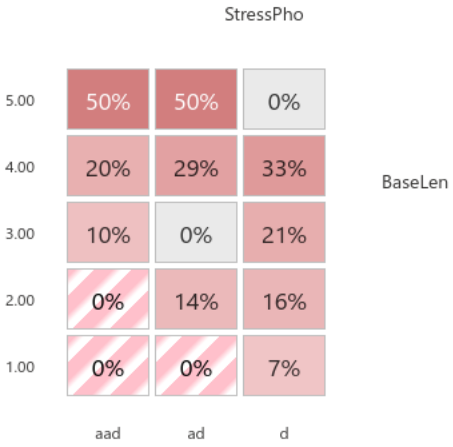

*Figure 2. Error rate for the length of the base noun in syllables and phonological stress position*

More information can be found on the left branch. It can be divided into two subbranches and concerns errors for the nouns based on their phonological stress position.

The out-most left subbranch concerns errors that occur in case when the length of base noun in syllables is less than 3.5 and more than 1.5 (i.e. 2 and 3 syllabic nouns), combined with the last stressed syllable property. The hierarchical error pattern here shows that the error rate for this particular combination of properties is higher than the average: 17.46%, 11 out of 63 nouns were misclassified. Among the misclassified nouns we encounter six 2-syllabic nouns (*glav(a)* 'leader', *dekabr'* 'December', *latyn'* 'Latin', *sentjabr'* 'September', *senat* 'senate', *raspad* 'disintegration') and five 3-syllabic nouns (*kardinal* 'cardinal', *xoxlom(a)* 'khokhloma (painting)', *seminar* 'seminar', *komitet* 'committee', *monastyr'* 'monastery').

The right subbranch of the tree is less interesting, since less errors can be found here. The error rate is 14.29% which is slightly lower than the overall error rate, only 6 out of 42 selected nouns were incorrectly classified (monosyllabic and 2-syllabic nouns where any syllable is stressed except for the last one). We will not focus on these error subset and analyze two previous subsets in more details.

The error heat map shown on Figure 2 allows to further investigate how the phonological properties in question impact the error rate across data subsets. Indeed, the highest error rates (up to 50%) are encountered for 5-syllabic base nouns, regardless phonological stress position. This heat map reveals that the error rates are also visibly higher for the nouns where the last syllable is stressed.

In previous section we assessed properties of base nouns which are statistically significant for suffix choice. Both the length of base noun in syllables and phonological stress position were not listed among these properties. Hence error analysis suggests that based one these properties we can isolate subsets of data with the highest error rates. But does this mean that these features are correlated to model errors?

Interpretability dashboard allows the exploration of the top important features that impact the overall model predictions. In previous section we saw that animacy, the last phoneme of the stem and the inflectional class are statistically significant in predicting if the suffix is *-n-* or *-sk-*. Not surprisingly, the visualizations available within Responsible AI toolbox prove the same, as shown on Figure 3 (`All data`). Moreover, it is possible to compare feature importance values for different selected subgroups of data side by side, for instance, the subgroups with the highest error rates (`BaseLenStressPho`: 2 and 3 syllabic nouns where the last syllable is stressed; `BaseLen`: 4 and 5 syllabic nouns).

Based on the information on feature importance and the ordering we can conclude that, in general, the model behaves in the same way on the whole data set and the two subgroups with highest errors (the only difference concerns 4 and 5 syllabic nouns: inflectional class appears to be slightly more important than the last phoneme for this data subset). This means that the same base nouns features are leveraged for predicting suffix across the three sets and that phonological stress position as well as the length of base noun in syllables are useful to isolate the majority of model errors, but they are not necessary correlated to these errors.

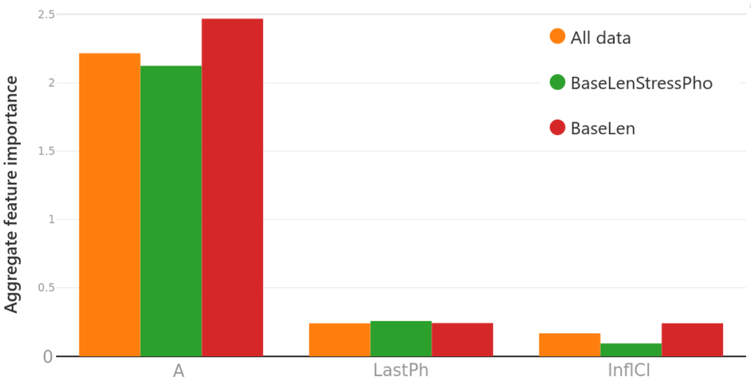

*Figure 3. Top 3 features by their importance*

In order to understand the reasons behind the erroneous predictions in test set we will contrast them to train data and to correctly classified data. For consistency, we will isolate the same subgroups in train subset and correct predictions as for incorrect predictions: 2 and 3 syllabic nouns with the last stressed syllable; 4 and 5 syllabic nouns.

| subset | CH | CC | PNH | CA | CH | CC | PNH | CA |
|---|---|---|---|---|---|---|---|---|
| **train: *-n-*** | 9 | 42 | 10 | 113 | 2 | 2 | 0 | 43 |
| **test: correct *-n-*** | 0 | 15 | 0 | 25 | 0 | 1 | 0 | 13 |
| **test: incorrect *-n-*** | 2 | 0 | 0 | 0 | 4 | 0 | 0 | 0 |
| **train: *-sk-*** | 55 | 6 | 43 | 16 | 20 | 2 | 47 | 10 |
| **test: correct *-sk-*** | 7 | 0 | 5 | 0 | 3 | 0 | 9 | 0 |
| **test: incorrect *-sk-*** | 0 | 1 | 0 | 8 | 0 | 2 | 0 | 4 |

*Table 6. Distribution of animacy across subsets:*
*2- and 3-syllabic nouns, the last stressed syllable | 4- and 5-syllabic nouns*
*(CH: ComHum, CC: ComConc, PNH: PropNHum, CA: ComAbst)*

Table 6 presents the distribution of animacy across two subset: 2- and 3-syllabic nouns with the last stressed syllable - in the left part of the table; 4- and 5-syllabic nouns - in the right part. Three main trends are observed here. First, the distributions of the most important base noun property to the suffix choice - animacy - are similar between 2-3-syllabic nouns with the last stressed syllable and 4-5-syllabic nouns. For instance, common abstract nouns are more numerous in both subsets for *-n-* training data (113 and 43 cases respectively). We observe the same tendencies in training set for *-sk-*: common human and proper non human nouns are the most represented (55 and 43 cases for the first subset and 20 and 47 - for the second). Second, train data distributions and correctly predicted data distributions follow the same patterns as well (common abstract nouns are the ones that are most numerous for *-n-* classification - 25 and 13 respectively in both subsets; similarly to train set, common human and proper non human nouns correctly predicted are the most numerous for *-sk-* (7 and 5; 3 and 9)). The third observation concerns test set where animacy has distinct distributions between correctly and incorrectly predicted data. For instance, if we take into consideration *-n-* distribution, we can see that common concrete nouns and common abstract nouns were correctly predicted with *-n-* suffix, whereas common human nouns (2 and 4 in both subsets) were mistakenly associated with *-sk-*. Similarly, with *-sk-* distribution, common human and proper non-human nouns are correctly identified with *-sk-*, but some common concrete (1 and 2) and common abstract nouns (8 and 4) were mistakenly classified with *-n-*.

The error cases are the following:

1. 2- and 3-syllabic nouns, the last stressed syllable
    - actual *-n-* suffix

    - `ComHum`: *glav*(*a*) 'leader', *kardinal* 'cardinal'
  - actual *-sk-* suffix
    - `ComConc`: *monastyr'* 'monastery'
    - `ComAbst`: *dekabr'* 'December', *komitet* committee', *latyn'* 'Latin', *raspad* 'disintegration', *seminar* 'seminar', *senat* 'senate', *sentjabr'* 'September', *xoxlom*(*a*) 'khokhloma (painting)'

2. 4- and 5-syllabic nouns
   - actual *-n-* suffix
     - `ComHum`: *gumanitarij* 'humanitarian', *professional* 'professional', *veterinar* 'vet', *čudovišč*(*e*) 'monster'
   - actual *-sk-* suffix
     - `ComConc`: *distrib'jutor* 'distributor', *universitet* 'university'
     - `ComAbst`: *artillerij*(*a*) 'infantry', *bogoslovi*(*e*) 'theology', *vselennaj*(*a*) 'universe', *žurnalistik*(*a*) 'journalism'

Even if the examples of common error patterns are not numerous, the conclusion is that misclassified data follows in general the distribution which is the opposite to the true suffix label. This can explain model errors: the model fails to discriminate correctly between two rival suffixes if the distribution of base noun properties is unusual (compared to the training data) for a specific suffix.

## 6. Conclusion

A brief literature overview given in Section 2 suggests that the topic of affix rivalry in denominal adjective formation in Russian is mostly approached with descriptive methods, statistical studies performed on a big corpus are missing. The modelization performed in Section 4 confirms the conclusions encountered in literature; in addition, it provides evidence on statistical significance of the properties of base nouns that allow to discriminate between the rival suffixes. Moreover, the error analysis performed in Section 5 sheds light on specific combination of properties that may behave differently and have a specific preference for the suffix which can't be drawn from the model.

This study was made using a logistic regression classifier in order to discriminate between *-n-* and *-sk-* adjectival suffixes in Russian. Overall, the model performs very well, with AUC ranging from 0.83 to 0.95, depending on train-test split. The choice of a simple logistic regression classifier is driven by its high transparency, since it allows an easy access to model parameters with feature importance and relevant statistics. For instance, the following base noun properties are statistically significant to predict *-n-* or *-sk-*: [+common,+human], [-common,+human] [+dental]-ending stems; to a lesser extent - [+labial]-, [+alveolar]-, and [+velar]-ending stems, inflectional class 2 and 1.

Compared to logistic regression, other classification models may not be interpretable that easily. Therefore, Responsible AI tools contribute to a better understanding of the output of ”black box” models. Even if logistic regression is transparent, it is nevertheless possible to get extra insights for this model through error analysis, and Responsible AI provides dashboards for relevant visual explorations which are easily interpretable as well.

The main tool used for the present study is binary tree which allows to isolate subsets of test data with the highest error rates. This complements the information about the most relevant features for the classification task with information on features that group data into subsets where model fails more often than on average. The overall error rate of the model is 14.71%, however, *-n-* and *-sk-* data may be grouped into subsets where error rates are even higher based on the length of the base noun in syllables and phonological stress position: 2 and 3 syllabic nouns with the last stressed syllable (11 nouns misclassified, error rate 17.46%), 4 and 5 syllabic nouns (10 nouns misclassified, error rate: 27.78%). These two subsets group data with more than two thirds of all misclassified nouns (11 false positives for *-sk-* and 19 false positives for *-n-*).

However, if it is possible to isolate error cases by certain phonological patterns, it does not necessarily implies that these exact patterns cause model failure. A closer look on aggregate feature importance suggests that the same properties are important for subclasses with the highest error rate and the whole data set. For instance, the most statistically significant property of the base noun that contribute to the suffix choice is animacy, and it remains significant across all the studied data sets (all data and two data sets with highest errors). The model failure can be explained by some cases of base noun properties distributions which do not follow the same patterns as in training set.

One possible extension of this approach would be including the combination of properties which lead to higher error rates as interaction terms in our model and to test weather it improves overall accuracies of the model and decreases the error rate. The approach used in this study should also be extended to additional binary classification problems (*-n-*/*-Ov-* and *-sk-*/*-Ov-*) and it may be applied to a multiclass classification involving all the three suffixes. This could provide a finer-grained quantitative evidence and potentially complete the discussion on suffix rivalry for denominal adjectives in Russian.

## Bibliography

Amershi, Saleema, Andrew Begel, Christian Bird, Robert DeLIne, Harald Gall, Ece Kamar, Nachi Nagappan, Besmira Nushi, and Tom Zimmermann. Software Engineering for Machine Learning: A Case Study. In *International Conference on Software Engineering (ICSE 2019) - Software Engineering in Practice track*. IEEE Computer Society, 2019. doi: 10.1109/ICSE-SEIP.2019.00042.

Arndt-Lappe, Sabine. Analogy in suffix rivalry: The case of English-ity and-ness. *English Language & Linguistics*, 18(3):497–548, 2014. doi: 10.1017/S136067431400015X.

Aronoff, Mark. Competition and the lexicon. In Elia, Annibale, Iacobini, Claudio, and Voghera, Miriam, editors, *Livelli di Analisi e fenomeni di interfaccia. Atti del XLVII congresso internazionale della Società Linguistica Italiana*, pages 39–52. Bulzoni, 2016.

Baayen, Harald, Anna Endresen, Laura A Janda, Anastasia Makarova, and Tore Nesset. Making choices in Russian: Pros and cons of statistical methods for rival forms. *Russian linguistics*, 37(3):253–291, 2013. doi: 10.1007/s11185-013-9118-6.

Baeskow, Heike. -ness and-ity: Phonological exponents of n or meaningful nominalizers of different adjectival domains? *Journal of English Linguistics*, 40(1):6–40, 2012. doi: 10.1177/0075424211405156.

Bansal, Gagan, Besmira Nushi, Ece Kamar, Dan Weld, Walter Lasecki, and Eric Horvitz. Updates in Human-AI Teams: Understanding and Addressing the Performance/Compatibility Tradeoff. In *AAAI Conference on Artificial Intelligence*. AAAI, 2019. doi: 10.1609/aaai.v33i01.33012429.

Bobkova, Natalia. Statistical modelization of suffixal rivalry in Russian: adjectival formations in -sk- and -n-. *Corpus*, (23), 2022. doi: 10.4000/corpus.6580.

Bobkova, Natalia and Fabio Montermini. Suffix rivalry in Russian: what low frequency words tell us. In *Mediterranean Morphology Meetings*, volume 12, pages 1–17, 2019.

Bonami, Olivier and Juliette Thuilier. A statistical approach to rivalry in lexeme formation: French *-iser* and *-ifier*. *Word Structure*, 11(2), 2018.

Chapman, Don and Royal Skousen. Analogical modeling and morphological change: the case of the adjectival negative prefix in English. *English Language & Linguistics*, 9(2):333–357, 2005. doi: 10.1017/S136067430500167X.

Fradin, Bernard. L'interprétation des nominalisations en N-age et N-ment en français. In Rainer, Franz, Russo, Michela, and Sanchez Miret, Fernando, editors, *Actes du XXVIIe congrès international de linguistique et philologie romanes* (*Nancy, 15-20 juillet 2013*). Société de linguistique romane/Eliphi, 2016.

Graudina, Ljudmila, Viktor Ickovič, and Lija Katlinskaja. *Grammatičeskaja pravil'nost' russkoj reči: stilističeskij slovar' varintov*. Nauka, 2001.

Graščenkov, Pavel. *Grammatika prilagatel'nogo. Tipologija ad"ektivnosti i atributivnosti*. Litres, 2019.

Guzmán Naranjo, Matías. *Analogical classification in formal grammar*. Language Science Press, 2019.

Guzmán Naranjo, Matías. Analogy, complexity and predictability in the Russian nominal inflection system. *Morphology*, 30(3):219–262, 2020. doi: 10.1007/s11525-020-09367-1.

Guzmán Naranjo, Matías and Olivier Bonami. Comparing derivational processes with distributional semantics. *ParadigMo II*, page 25, 2021.

Hathout, Nabil. Une approche topologique de la construction des mots: propositions théoriques et application à la préfixation en anti. *Des unités morphologiques au lexique*, pages 251–318, 2011.

Huyghe, Richard and Marine Wauquier. Distributional semantics insights on agentive suffix rivalry in French. *Word Structure*, 14(3):354–391, 2021. doi: 10.3366/word.2021.0194.

Hénault, Christine and Sergueï Sakhno. Çem supermarket-n-yj luçşe supermarket-sk-ogo? Slovoobrazovatel'naja sinonimija v russkix ad"ektivnyj neologizmax po dannym interneta. *B. Tošovic, A. Wonisch. Wortbildung und Internet*, 2015.

Kapatsinski, Vsevolod. Velar palatalization in Russian and artificial grammar: Constraints on models of morphophonology. *Laboratory phonology*, 1(2):361–393, 2010. doi: 10.1515/labphon.2010.019.

King, David, Andrea Sims, and Micha Elsner. Interpreting sequence-to-sequence models for Russian inflectional morphology. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 409–418, 2020.

Kustova. Prilagatel'nye. *Materialy k korpusnoj grammatike russkogo jazyka. Vyp.3. Časti reči i leksiko-grammatičeskie klassy*, pages 40–107, 2018.

Lignon, Stéphanie. –iser and –ifier suffixations in French: Verify data to verize hypotheses? In *Décembrettes 7*, 2010.

Lindsay, Mark and Mark Aronoff. Natural selection in self-organizing morphological systems. In *Morphology in Toulouse. Selected Proceedings of Décembrettes 7* (*Toulouse 2-3 December 2010*), pages 133–153. Lincom Europa, 2013.

Missud, Alice and Florence Villoing. The morphology of rival-ion,-age and-ment selected verbal bases. *Dany Amiot Delphine Tribout*, page 29, 2020.

Nushi, Besmira, Ece Kamar, and Eric Horvitz. Towards Accountable AI: Hybrid Human-Machine Analyses for Characterizing System Failure. In *HCOMP 2018*. AAAI, 2018.

Parker, Jeff and Andrea Sims. Irregularity, paradigmatic layers, and the complexity of inflection class systems: A study of Russian nouns. *InP. Arkadiev & F. Gardani Eds. The Complexities of Morphology*, 2019. doi: 10.1093/oso/9780198861287.003.0002.

Plénat, Marc. Enquête sur divers effets des contraintes dissimilatives en français. In Roché, Michel, Boyé, Gilles, Hathout, Nabil, Lignon, Stéphanie, and Plénat, Mark, editors, *Des unités morphologiques au lexique. Paris: Hermès-Lavoisier*, pages 145–190, 2011.

Plungjan, Vladimir, Tat'jana Reznikova, and Dmitrij Sičinava. Nacional'nyj korpus russogo jazyka: obščaja xarakteristika. *Naučno-texničeskaja informacija. Serija 2: Informacionnye processy i sistemy*, (3):9–13, 2005.

Roché, Michel. Quel traitement unifié pour les dérivations en-isme et en-iste? In Roché, Michel, Boyé, Gilles, Hathout, Nabil, Lignon, Stéphanie, and Plénat, Mark, editors, *Des unités morphologiques au lexique. Paris: Hermès-Lavoisier*, pages 69–143. Hermès-Lavoisier, 2011.

Seabold, Skipper and Josef Perktold. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010. doi: 10.25080/Majora-92bf1922-011.

Sims, Andrea. Slavic morphology: Recent approaches to classic problems, illustrated with Russian. *Journal of Slavic Linguistics*, 25(2):489–524, 2017. doi: 10.1353/jsl.2017.0019.

Srivastava, Megha, Besmira Nushi, Ece Kamar, Shital Shah, and Eric Horvitz. An Empirical Analysis of Backward Compatibility in Machine Learning Systems. In *KDD*, 2020. doi: 10.1145/3394486.3403379.

Švedova, Natal'ja. *Russkaja grammatika*, volume 1. Moskva: Nauka, 1980.

Thuilier, Juliette. *Contraintes préférentielles et ordre des mots en français*. PhD thesis, Université Paris-Diderot-Paris VII, 2012.

Timberlake, Alan. *A reference grammar of Russian*. Cambridge University Press, 2004.

Townsend, Charles Edward. *Russian word-formation*. Slavica Publishers, 1975.

Varvara, Rossella. Constraints on nominalizations: Investigating the productivity domain of Italian-mento and-zione. *Zeitschrift für Wortbildung/Journal of Word Formation*, 4(2):78–99, 2020. doi: 10.3726/zwjw.2020.02.05.

Wauquier, Marine. *Confrontation des procédés dérivationnels et des catégories sémantiques dans les modèles distributionnels*. PhD thesis, Université Toulouse II Jean Jaurès, 2020.

Zaliznjak, Andrej. *Grammatičeskij slovar' russkogo jazyka*. Russkie slovari, 2003.

Zemskaya, Elena. *Jazyk kak dejatel'nost'. Morfema, slovo, reč*. Moskva: Flinta, 2015.

**Address for correspondence:**
Natalia Bobkova
natalia.bobkova@univ-tlse2.fr
Université de Toulouse 2 Jean Jaurès, Maison de la recherche, B503
5, allée Antonio Machado 31058 Toulouse cedex 9, France

PBML

The Prague Bulletin of Mathematical Linguistics
NUMBER 118 APRIL 2022 25–53

# Hebrewnette – A New Derivational Resource for Non-concatenative Morphology: Principles, Design and Implementation

Lior Laks,[a] Fiammetta Namer[b]

[a] Bar-Ilan University – Ramat-Gan, Israel
[b] UMR 7118 ATILF, Université de Lorraine & CNRS – Nancy, France

## Abstract

This paper presents the architecture of a derivational database of Modern Hebrew (and more generally of Semitic languages) called Hebrewnette. The methodology adopted is based on adjusting the structure and properties of a database developed for the description of the derivational relations in the lexicon of a Romance language (Démonette), and providing it with additional features to account for the specificities of the morphology of Semitic languages, with special reference to root-and-pattern non-concatenative morphology. We present the properties of Hebrewnette and the type of information it consists of, with emphasis on both structural and semantic relations between words. We show how this is implemented and examine two case studies, where we demonstrate how the annotations that are used allow us to verify theoretical hypotheses about non-concatenative morphology. The design of Démonette's annotation system allow its features, initially designed for French, to capture morphological and semantic relations between Hebrew words, regardless of the type of morphology (concatenative or non-concatenative).

## 1. Introduction

This paper presents a derivational resource for Modern Hebrew based on an existing infrastructure that was originally designed for Romance languages like French. We show how the existing architecture of the database can be adapted for Semitic morphology with some relevant additions.

Corresponding author: fiammetta.namer@univ-lorraine.fr
Cite as: Lior Laks, Fiammetta Namer. Hebrewnette – A New Derivational Resource for Non-concatenative Morphology: Principles, Design and Implementation . The Prague Bulletin of Mathematical Linguistics No. 118, 2022, pp. 25–53. doi: 10.14712/00326585.018.

Current available resources and tools for European languages can be divided in two main types:

- The first ones specifically describe a process (or a family of processes) of a given language. The reader can refer to Kyjánek (2018) for a typological description of the structure and coverage of 30 recent derivational resources for Romance (including Latin), Germanic and Slavic languages, which provides a complete list of the main existing derivational databases and lexicons with derivational annotations.
- The second type of databases aims at a multilingual description by homogenizing linguistically and structurally heterogeneous sources: construction of standards in terms of tagsets (McCarthy et al. 2018), standardization of existing resources in the form of architectures with a universal vocation (Universal Derivation (Kyjánek et al. 2020), MorphyNet (Batsuren et al. 2021), UniMorph (Kirov et al. 2018)). This second category of databases is fed (among other things) by the content of the first ones.

This article is devoted to the first of these two categories of databases. Specifically, we ask how a database designed and developed to represent the derivational properties of French - and more broadly, of a Romance language - can be used to describe the morphology of Semitic languages, and more particularly the non-concatenative derivational relations. The database we use as a starting point for this study is Démonette (Namer and Hathout 2020). In order to address this issue, we break it down into three questions.

- What is needed, in the design of a database, to represent in a satisfactory, exhaustive and fine-grained way the derivational properties of the non-concatenative (as well as concatenative) morphology of Semitic languages?
- Since Démonette is designed for the fundamentally concatenative morphology of a language like French, thus genotypically very distant, could its principles be applied for this purpose, by means of additional descriptions but without modifying the existing structure? If yes, this can serve as an initial prototype for a unified framework for the description of the derivational morphology of many languages.
- Does the database succeed in representing theoretical issues regarding Semitic Morphology?

To answer these three questions, our paper is structured as follows. In Section 2, we recall the morphological principles that distinguish Hebrew (and Semitic languages in general) from Western European languages. Then, we present Démonette, the derivational database of French that we adapt to Hebrew (Section 3). Section 4 describes the extensions (new attributes, new values) that formalize the specific derivational properties of Hebrew in the Hebrewnette database, while preserving the architecture of the source database and keeping the original features and values. Section 5 presents two case studies, which put the Hebrewnette feature structure to the test.

Finally, in Section 6, we explain how we chose the entries in order to build the current version of Hebrewnette, with the aim of evaluating the relevance of the proposed tagset, and assessing its capability of covering a wide variety of types of structural and semantic relations in Hebrew.

## 2. Hebrew Morphology

### 2.1. Root-and-pattern Morphology

Hebrew word formation relies highly on non-concatenative morphology, i.e. via root and pattern (Aronoff 1994, Berman 1978, Bolozky 1978, Ravid 1990, Schwarzwald 1981). The pattern indicates the prosodic structure of the word and it consists of the following elements: (i) consonantal slots; (ii) vocalic pattern; and in some cases (iii) affixes (Bat-El 1994, 2017). For example, the verbs *siper* 'tell$_V$' and *limed* 'teach$_V$' are formed in the CiCeC pattern.[1] They share the vocalic pattern i-e and differentiate in their roots, s-p-r and l-m-d respectively. The verbs *hitkavec* 'become shrunk$_V$' and *hitraxev* 'become wide$_V$' are formed in the hitCaCeC pattern, which consists of the prefix *hit-*, in addition to the vocalic pattern a-e.

#### 2.1.1. Verb patterns

Words that share the same consonantal root typically share some semantic relations with different degrees of transparency, for example *hidpis* (hiCCiC) 'print$_V$', *hudpas* (huCCaC) 'be printed$_V$', *madpeset* (maCCeCet) 'printer$_N$' and *tadpis* (taCCiC) 'printout$_N$'. Hebrew verbal patterns typically differ from each other with respect to transitivity and the semantic types of verbs that they host (see Aronoff 1994, Berman 1978, Bolozky 1978, Borer 1991, Doron 2003, Ravid 1990, Ravid et al. 2016: and references therein). For example, CiCeC typically hosts active transitive verbs, e.g. *kivec* 'shrink', *nigev* 'wipe' and *xibek* 'hug', while hitCaCeC typically hosts intransitive verbs like inchoatives (*hitkavec* 'become shrunk'), reflexives (*hitnagev* 'wipe oneself') and reciprocals (*hitxabek* 'hug each other'). However, these only represent tendencies and there is no one-to-one correspondence between form and meaning of the patterns. For example, *hitpalel* 'pray' is formed in hitCaCeC but does not belong to any of the above mentioned semantic classes.

Within verb formation, non-concatenative formation is obligatory, and every verb that enters the language must conform to one of the existing patterns. This is attested in the formation of verbs that are derived from words without roots, including words borrowed from various languages. For example, the verbs *midel* 'make a model (out

[1] The term "formed" indicates that a specific word shares the form of one of the existing patterns. It does not necessarily imply that a word formation pattern actually took place. Rather, it denotes the fact that (a) the word has the vocalic melody and the affixes (if any) of that pattern, and (b) the root consonants of this word occupy the consonantal slots of the pattern.

of smth)' (CiCeC) , *hispim* 'send a spam' (hiCCiC) and *hitfakes* 'be in focus' (hitCaCeC) are derived from English loan words and are formed in three different patterns. Most transitive verbs are formed in CiCeC by default and some are formed in hiCCiC in order to preserve the consonant cluster of the base (*spam-hispim*). The verb *hitfakes* is formed in hitCaCeC because it is an intransitive verb. Newly coined intransitive verbs that are inchoative, reflexive and reciprocal are formed in hitCaCeC almost exclusively.

### 2.1.2. Nominal and adjectival patterns

Hebrew also has a set of patterns that are used for the formation of nouns and adjectives. Most patterns have typical meanings, although meanings that are associated to them represent mainly tendencies. For example, the maCCeC pattern is typical of instrument nouns like *masnen* 'filter' and *masrek* 'comb', but is used for the formation of other nouns, e.g. *martef* 'basement'. The CaCaC pattern is typical of agent nouns (e.g. *cayar* 'painter'), but is also used for instrument nouns (e.g. *vasat* 'regulator') and adjectives (e.g. *raxav* 'wide').

Each verbal pattern, apart from the passive patterns CuCaC and huCCaC, is related to a typical nominal pattern that is used for the formation of actions nouns. For example, the typical action noun pattern of CaCaC is CCiCa (e.g. *katav* 'write' - *ktiva* 'writing') and that of CiCeC is CiCuC (e.g. *limed* 'teach' - *limud* 'teaching'). There is some interpredictibility between the verbal and nominal patterns, and this allows us to feed the database in a semi-automatic way (see Section 6). However, this system is subject to a certian amount of irregularity. For example, the action noun of *higer* (CiCeC) 'emigrate' is *hagira* and not **higur*. Moreover, some of the action nouns have an additional nominal meaning. This is a well-known action-result polysemy phenomenon, where the deverbal action noun also denotes the result of such action. Such polysemy can be found in many languages, as has been shown in various studies (see, among others, Alexiadou 2001, Berman and Seroussi 2011, Borer 2014, Comrie and Thompson 2007, Grimshaw 1990, Hazout 1995, Levin and Rappaport Hovav 2005, Melloni 2011, Ravid and Avidor 1998: and references therein). For example, the action noun of the CiCeC verb *pirsem* 'publish' is *pirsum* (CiCuC), which denotes both the action of publishing and the noun 'publication'.

Each verb pattern has a participle pattern that is used to indicate present tense of verbs. Participle patterns are polycategorial, as they are also used to denote nouns and adjectives. For example, CoCeC is the participle pattern that is related to the CaCaC verbal pattern. For example, *lomed* denotes both the participle form of the verb *lamad* 'learn' and the agent noun 'learner'. The participle form *meratek* (meCaCeC) denotes both the participle form of the CiCeC verb *ritek* 'fascinate' and the adjective 'fascinating'. In addition, some words are formed in participle patterns and do not have verbal counterparts. For example, the agent noun *šoter* 'policeman' is formed in the CoCeC pattern and there is no CaCaC verb like **šatar*. Participle patterns are in general multifunctional, as can be attested in other languages as well (for instance, verb-to-adjective conversion in French according to Tribout (2010)).

There is a special group of disyllabic patterns called Segolates: CéCeC, CáCaC, CóCeC and CóCaC. They differ from other patterns in three main aspects (see Bat-El 2012, Bolozky 1995, Schwarzwald 2002: among others). First, while most words that are formed in patterns have ultimate stress (e.g. *masrék* 'a comb') the Segolate patterns always have penultimate stress, e.g. CéCeC (*kéšer* and CóCeC (*tóxen* 'content').[2] Second, they are not associated with typical meanings and host a variety of nouns. Third, their inflectional paradigms exhibit three surface stems. For example, *kéšer* 'relation', *kšar-im* 'relations' and *kišr-i* 'my relation'.

### 2.1.3. Root types and relations between roots

Most roots consist of 3 consonants, but four-consonant roots are also present in the lexicon in a non-negligible proportion. These are found almost exclusively in the CiCeC, CuCaC and hitCaCeC patterns whose prosodic structure can accommodate more than 3 consonants, e.g. p-rs-m for *pirsem* (CiCCeC) 'publish'. Some roots are weak in the sense that one or more of the consonants do not surface in all forms or do not surface at all. For example, the root of the verb *rac* (CaCaC) 'run' is r-W-c, where the W never surfaces and can only be associated with the verb through diachronic analysis. In addition, some phones undergo phonological alternation in the transition between patterns, e.g. stop-fricative alternations, as in *gavar* (CaCaC) 'increase$_{inchoative}$' - *higbir* (hiCCiC) 'increase$_{transitive}$'.

By default, a relation connects two or more items that share the same root. This is one of the main features of Semitic morphology that is responsible for a rich system of derivational paradigms that revolve around the consonantal root. However, there are relations that connect items with different roots. These particular relations surface in cases where a consonant is added to the root. This type of relation creates a new family, and its members share the new root. The two families form different paradigms. Let us demonstrate it with respect to the pair *tadrix* 'briefing$_N$' - *tidrex* 'debrief$_V$'. The taCCiC pattern, which includes the prefix *ta-*, is used for the formation of different nouns that can be related to verbs in different patterns, e.g., *hidpis* (hiCCiC) 'print$_V$' - *tadpis* 'printout$_N$'. The noun *tadrix* 'briefing' is formed in the taCCiC pattern, and is semantically related to the hiCCiC verb *hidrix* 'guide$_V$' and the haCCaCa action noun *hadraxa* 'guidance$_N$'. The three words form a derivational family sharing the consonantal root d-r-x. The verb *tidrex* 'debrief$_V$' is formed in the CiCeC pattern based on the noun *tadrix*, taking the t consonant of the derivational prefix *ta-* as part of the new root t-dr-x. The CiCeC pattern is paradigmatically connected to the CiCuC pattern of action nouns (*tidrux* 'debriefing$_N$') and to the passive CuCaC pattern (*tudrax*

[2] When words and patterns have penultimate stress, this is marked throughout the paper by an acute accent. Otherwise, patterns and words are left unmarked.

'be debriefed$_V$'). The CiCeC pattern of *tidrex* induces new types of relations within its new family.

### 2.2. Other word formation strategies

In contrast to verbs, the formation of nouns and adjectives is based on a variety of word formation strategies, most of which are highly productive in European languages. Nouns, for example, can be 'raw'[3], (*daf* 'page'), borrowed (*lazanya* 'lasagna'), and can be formed via different word formation processes. Hebrew has a set of derivational affixes that are used for the formation of nouns and adjectives. Affixes can be attached to different stems with or without a morphological structure. They can be attached to raw stems. For example, the noun *yam* 'sea' takes the suffixes *-i* and *-ay* to derive the adjective *yam-i* 'marine' and the agent noun *yam-ay* 'sailor'. The adjective *kal* 'easy' takes the suffix *-ut* to derive the abstract noun *kal-ut* 'easiness'. Affixes are also attached to words with root and pattern. For example, the agent noun *nagar* 'carpenter' is formed in the CaCaC pattern, and the suffix *-iya* is attached to form the location noun *nagar-iya* 'carpentry shop'. Some words undergo morpho-phonological alternations when affixes are attached. For example, *šémeš* 'sun' undergoes two alternations in the formation of the adjective *šimš-i* 'sunny'; the first vowel changes from *e* to *i*, and the second vowel is deleted.

In addition to affixation, nouns and adjectives are formed by other word formation strategies like reduplication (*xatul* 'cat' - *xataltul* 'kitten'), acronym formation (e.g. *ramax* 'department chair', based on *roš* 'head' and *maxlaka* 'department'), blending (e.g. *midrexov* 'pedestrian mall', based on *midraxa* 'sidewalk' and *rexov* 'street') and compounding (e.g. *bet-sefer* 'school', lit. house-book).

### 2.3. Word-based approach to Semitic morphology

The design of Démonette (Section 3), and specifically its implementation into Hebrewnette, relies on word-based models to word formation. The word-based approach, originally proposed in (Aronoff 1976), assumes that the mental lexicon consists of actual words rather than morphemes, roots or coded concepts. Aronoff's main claim is that a word is formed by applying a Word Formation Rule (WFR) to an existing word or stem. They serve for producing and understanding new words, which may be added to the speaker's lexicon and as redundancy rules (Jackendoff 1975) defining morphological relations. Such a view assumes a phonological representation of words in the lexicon. The distinction between a root/morpheme-based morphology and a word-based morphology corresponds to the traditional distinction between 'item and arrangement' models and 'item and process' models respectively (Anderson 1992,

[3] we use 'raw' following Schwarzwald (2002), to indicate that a word has no complex morphological structure. It is not derived from another word, it is not formed in a pattern and does not consist of affixes.

Hockett 1954, Matthews 1972, 1974). In the former model, morphemes are the basic units of meaning and they are arranged linearly, while in the latter model, word structure is specified by a series of processes.

Semitic morphology raises questions about the exact processes that take place in word formation. We adopt the theory of Stem Modification (Bat-El 1994, 2017, 2019, McCarthy and Prince 1990, Steriade 1988), which accounts for generalizations about morpho-phonological alternations by allowing for stem-internal adjustments rather than positing the extraction of a consonantal root (Bat-El 1986, Davis and Zawaydeh 2001, Farwaneh 1990, Goldenberg 1994, Hoberman 1992, Idrissi and Kehayia 2004, McCarthy 1981, McCarthy and Prince 1986, Ornan 1983, Yip 1988: among others). Stem modification accounts for the transfer of information like prosodic structure from a base form to a derived form. It also provides a uniform account for morphological phenomena in non-Semitic languages, which are similar to those of non-concatenative morphology, e.g. ablaut in *sing/sang/song* (Bat-El 2002). Various studies have highlighted the absence of motivation for assuming an independent mechanism of root extraction (Aronoff 2007, Bat-El 1994, 2017, Benmamoun 2003, Bolozky 1999, 2012, Hammond 1988, Heath 1987, Kihm 2011, McCarthy and Prince 1990, Ratcliffe 1997, Rose 1998, Ussishkin 2005: among others). The status of the consonantal root is under an ongoing debate and there are different approaches with regard to its necessity and the actual mechanism that applies in word formation (Faust 2019, Nevins 2005, Rasin et al. 2021). It is important to emphasize that root-based approaches do not assume that Semitic word formation relies only on the representation of the consonantal root. Under such approaches, some words are derived directly from roots, while other words are derived directly from words (Arad 2005, Doron 2003, Faust 2015, Kastner 2019, 2020). Words that are derived from other words via non-concatenative morphology have to conform to one of the existing patterns. This is executed via “template imposition” (Faust and Hever 2010), where the pattern is imposed on the derived word based on its base. The question under debate is about the exact process that template/pattern imposition involves.

The design of Hebrewnette is based on a Stem Modification approach, as it represents, among other features, alternations that take place in the transition between words within paradigms. Such alternations relate both to the consonantal root and other parts of words. As will be detailed in Section 4, and demonstrated with respect to the cases studies in Section 5, the design of Hebrewnette provides the relevant information that is needed to examine structural relations between words which are formed in non-concatenative morphology (in addition to words formed by other processes), and such relations go beyond the consonantal root.

## 3. Démonette's principles

The founding principles of Démonette (Hathout and Namer 2016, Namer and Hathout 2020) that have been applied to Hebrewnette are the following (see also Laks and Namer 2020):

- Each entry describes a derivational relation between two lexemes, that is, unmarked words.
- Entries form derivational families represented by connected graphs, where derivational families are defined as sets of derivationally related lexemes (Hathout 2011).
- Lexemes and relations are described in two separate tables. The table of lexemes displays properties of words independent of the morphological relations these words can be involved in.
- Derivational relations occur for any pair of members of a given family. Relations are labelled according to their specific properties, as well as the properties (morphological, categorical, semantic, phonological) of the lexemes connected by the relation.
- These complex labels are the combination of several feature values. We exemplify them with the family of *banque* 'bank$_{\mathrm{N}}$', *banquier* 'banker$_{\mathrm{N}}$', *bancaire* 'of a bank$_{\mathrm{A}}$', *interbancaire* 'between banks$_{\mathrm{A}}$'.

Relations are distinguished according to their orientation, that is, whether one of the two connected lexemes is the ancestor of the other (in Table 1-a, `as2des` says that *banque* is the ancestor of *bancaire*; in Table 1-a', the reverse relation `des2as` indicates that the lexeme L1 *bancaire* is a descendant of L2 *banque*), or not (Table 1-b). Examples like Table 1-b include instances of cross-formation, where two co-derived words (like *prédateur* 'predator$_{\mathrm{N}mas}$' and *prédatrice* 'predator$_{\mathrm{N}fem}$', in French, Table 1-c) may lack a common ancestor (e.g. the verb **préder* is not attested). Orientation may be undecidable, and therefore labelled `NA`, as with the (*performant*$_{\mathrm{A}}$ 'performing', *performance*$_{\mathrm{N}}$ 'performance') conversion in Table 1-h.

When relevant, "ancestorhood" is evaluated based on semantic criteria. Let us examine the pair (*vivisecter*$_{\mathrm{V}}$ 'vivisect', *vivisection*$_{\mathrm{N}}$ 'vivisection'), Table 1-d. From a morphological (that is, formal) point of view, the noun seems to be derived from the verb by suffixation in *-ion* (as for example, *infection*$_{\mathrm{N}}$ 'infection' is derived from *infecter*$_{\mathrm{V}}$ 'infect'). However, the verb *vivisecter* is much more recent than *vivisection* and much less frequent (the Google search with the infinitive verb form results in approximately 2.000 hits, whereas it results in more than 2.5 millions with the singular noun form). Most importantly, unlike *infection*$_{\mathrm{N}}$ which is undoubtedly interpreted as the action noun of *infecter*$_{\mathrm{V}}$, *vivisection*$_{\mathrm{N}}$ cannot be defined with respect to the semantic content of the verb (*vivisection* is by no mean 'the action of vivisecter$_{\mathrm{V}}$'); on the contrary, the noun is the semantic base of the verb, which can be defined as 'to practice a *vivisection*$_{\mathrm{N}}$'. The orientation value `as2des` indicates that the noun is the ancestor

| | L1 | L2 | Orientation | Morpho. pattern1 | Morpho. pattern2 | Complexity |
|---|---|---|---|---|---|---|
| a | banque | bancaire | as2des | X | Xaire | simple |
| a′ | bancaire | banque | des2as | Xaire | X | simple |
| b | banquier | bancaire | indirect | Xier | Xaire | simple |
| e | bancaire | interbancaire | as2des | X | interX | motiv-form |
| f | banque | interbancaire | as2des | X | interXaire | motiv-sem |
| g | banquier | interbancaire | indirect | Xier | interXaire | complex |
| c | prédateur | prédatrice | indirect | Xeur | Xrice | simple |
| d | vivisection | vivisecter | as2des | Xion | X | simple |
| h | performant | performance | NA | X | X | simple |

*Table 1. Orientation and complexity of derivational relations in Démonette*

| | L1 | L2 | Pattern1 | Pattern2 | Formal variation | Complexity | Cross-definition |
|---|---|---|---|---|---|---|---|
| a | fleur<br>/flœʁ/ | fleurette<br>/flœʁɛt/ | X | Xette | NA | simple | a *fleurette* is a small *fleur* |
| b | fleur<br>/flœʁ/ | floral<br>/flɔʁal/ | X | Xal | /œ/ ~ /ɔ/ | simple | — |
| c | fleur<br>/flœʁ/ | anthophobe<br>/ãtofɔb/ | X | Xphobe | NA | motiv-sem | they who fear *fleur*s are *anthophobe*s |

*Table 2. Formal variation and cross-definitions in Démonette*

in the relation. Divergence between form and meaning, such as the verb formation in Table 1-d corresponds to the phenomenon of so-called back-formation (Becker 1993).

What makes rows (d) and (a) (for instance) in Table 1 two different cases is the combination of the orientation value with the morphological pattern of the two lexemes involved. As we can see, `X` is the pattern of the descendant of the relation in Table 1-d, whereas it is that of the ancestor in Table 1-a. Notice that the pattern of a lexeme depends on the relation where it occurs. For instance, in (*bancaire*, *banque*) Table 1-a, the pattern of *bancaire* is `Xaire`, and that of *banque* is `X`, where `X` represents the stem /bɑ̃k/ they share. In contrast, *bancaire* in Table 1-e is connected to *interbancaire*, and the shared stem is /bɑ̃kɛʁ/, therefore the pattern of *bancaire* is `X` and that of *interbancaire* is `interX`.

Another key feature of a derivational relation is its complexity. For regular relations, the value is `simple` when either one of the two lexemes is the base for the derivation of the other Table 1-a, or when both lexemes are daughters of the same base (even when this base is not or no more attested) as in Table 1-b,c. It is `complex` otherwise, as in Table 1-g. Not all derivational relations are morphologically canonical (in the sense of Corbett (2010)). They may involve form-meaning mismatches (Hathout and Namer 2014, Namer and Hathout 2020). This is the case for parasynthetic phenomena (Iacobini 2020), of which *interbancaire* is one of the many illustrations (Table 1-e,f). On the one hand, this adjective is formally derived from the adjective *bancaire* (Table 1-e). But on the other hand, its semantic content directly depends on that of the noun *banque*, since *interbancaire* means 'between several banks' and not 'between things related to banks'. So, semantically, *interbancaire* is derived from *banque*. This dual motivation is expressed by two new values of complexity: `motiv-form` indicates that the relation is uniquely motivated formally (but not semantically), and `motiv-sem` expresses direct interpretative filiation (but a lack of formal transparency).

Besides structural properties, a relation within a derivational family carries phonological features that describe the way the relation affects the stems of the related lexemes, as illustrated in Table 2 with examples of lexemes derivationally connected to the noun *fleur* 'flower'. There is formal identity when at least one of L1's stems is identical to one of L2's stems, as in Table 2-a. Otherwise, phonological variations are ranked according to morphophonological features. In Table 2-b, the only variation at play is an instance of vowel backness. For stems that are historically related but are unrelated from a synchronic perspective, there is no phonological variation encoded, but the value of complexity is `motiv-sem`. This is the case of *antho-* (/ɑ̃to/) in Table 2-c; this Greek learned suppletive component of the noun *fleur* /flœʁ/ 'flower' occurs almost only in neoclassical compounds.

Finally, relations are encoded with features describing their semantic properties. Based on the ontological class of each lexeme, these properties include the semantic category of the relation, eg., `agent-activity` for (*prédateur*, *prédation*), `location-agent`, for (*banque*, *banquier*), or `identity` for relations between words with the same semantic content e.g. (*banque*, *bancaire*) (in line with Spencer' (2013) notion of transposition). In addition, relations are described by means of a paraphrase cross-defining the related words, cf. Table 2-a,c last column.

To sum up, the database is deliberately designed as highly redundant. Each lexical unit has as many derivational descriptions as it has multiple relations within its family. In addition to the properties of its relation with other words, each lexical unit is defined by features independent of the relations in which it is found (e.g. its inflectional paradigm, part of speech, ontological category, frequency).

A morphological description is therefore the result of the interaction of formal, categorical and semantic properties. These three levels of description are autonomous, which allows us to represent non-canonical phenomena, such as derivations involving meaning-form discrepancies (back- & cross-formation, parasynthesis, conversion, bracketing paradoxes, etc. see Hathout and Namer (2014)) straightforwardly. At each level of description, properties are abstracted away into patterns that generalize the different sorts of regularities that can be found in the constructed lexicon: phonological, semantic, morphological. In other words, Démonette implements the principles of the paradigmatic approach to morphology (Bonami and Strnadová 2018, Hathout and Namer 2022): (i) relations where the same lexeme is involved combine two-by-two and form connected graphs that represent derivational families; (ii) these families can be superposed when the relations between their members instantiate the same abstract properties. In sum, the architecture is designed to integrate paradigmatic relations in morphology, which is also a characteristic of Hebrew, as we have seen in Section 2.

## 4. The Hebrewnette database: basic description

The design of Hebrewnette, based on the basic principles of Démonette's annotation system, makes its features suitable for capturing both morphological and semantic relations between Hebrew words, regardless of the type of morphology (i.e. concatenative or non-concatenative). Other morphological tools and resources for Hebrew and other Semitic languages exist: see, for example, Daya et al. (2008), Itai and Wintner (2008), Klimek et al. (2016), Neme (2011), Nir et al. (2013), Singh and Habash (2012), Wintner (2004). However, they rely mostly on the consonantal root as the central entity used as a base for word formation, which implies that family networks are oriented tree-shaped graphs, where only ancestor-descendant relations are represented, and not paradigmatic relations between words. The design of Hebrewnette relies on a word-based approach to morphology, and it therefore allows a separation between structural and semantic properties, in the analysis of such paradigmatic relations. Note, however, that the properties of Hebrewnette can be used to analyse Hebrew morphology under both word-based and root-based approaches.

Given the non-concatenative nature of the morphology of Semitic languages (Section 2) and the structures already present in Démonette (Section 3), a number of extensions are necessary when transposing Démonette for the analysis of derivation in

Hebrew. In a nutshell, these extensions regard the description of (relations between) patterns and (relations between) roots (see Section 2.1.3). Other typical features of Hebrew, e.g. those presented in Section 2.1.1, are in line with the predictability of lexical semantic properties (argument structure, agentivity) of words whose pattern belongs to the same derivational paradigm.

First, new attributes are needed to describe the **internal (morphological) structure** and the **root** of the lexemes connected by a derivational relation:

- We have seen (Section 2.1.1) that verbs have root and pattern, while nouns and adjectives can be formed both by root and pattern (Section 2.1.2) and by other word formation processes (Section 2.2). The pattern is indicated for words that are formed via non-concatenative morphology, e.g. `CiCeC` and `CaCaC`, Table 3-a,b. Some patterns include affixes, e.g. `hiCCiC`, Table 3-c, and `hitCaCeC`, Table 3-d. In addition, derivational affixes can be attached to words that have root and pattern `CaCaC+iya`, Table 3-e.
- Some words do not have root and pattern (Section 2.2). In that case, their morphological structure is coded as `raw`, Table 3-f, or `borrowed`, Table 3-g. The pattern of words that are formed in Segolate patterns (Section 2.1.2) are marked with an accent (Table 3-h,l).
- For words with a pattern, the morphological structure is displayed in the form of its vowel schema between “|” (`|ie|`, Table 3-a), completed when relevant with pattern affixes (`hi|0e|`, Table 3-c, where `0` indicates an empty vowel position, `hit|ae|`, Table 3-d). In case of affixation, affixes which are not part of the pattern are separated from the base structure by ‘+’ (`|aa|+iya`, Table 3-e). This feature is significant also for borrowed words (Table 3-c), when they are the source of new patterns, cf. Table 4-c.
- When relevant, the formal variation between a word and its pattern is explicitly indicated: for example, Table 3-i, *rac* is an instantiation of the `CaCaC` pattern, where the second consonant of the pattern is missing `C2:0`, as well as its second vowel `V2:a~0` (see Section 2.1.3).
- Raw and borrowed words have no root (`NA` stands for ‘not applicable’, Table 3-f,g). For other words, roots are classified according to their type. Three-consonant roots are labelled `r(egular)`, e.g. Table 3-a,b. In four-consonant roots (`r=4`) the middle consonant position is instantiated by a cluster made of the second and third consonants (`p-rs-m`, Table 3-j). Phonological or orthographical identity between regular roots is expressed by a specific value. The value `r-hom` is used in case of homonymy. For instance, Table 3-k, the root `s-p-r` is used in two semantically unrelated derivational families: one containing *sipur*$_\text{N}$ ‘story’, and one *sapar*$_\text{N}$ ‘barber’ and both containing the ambiguous verb *siper*, which denotes either ‘tell’ or ‘cut hair’ (Table 3-k). The value `r-hoph` indicates a case of homophony. For example, the consonant /k/ of the phonological root `k-š-r` corresponds to two different spellings: ך (כשר) for the family of *kóšer*$_\text{N}$ ‘ability’ (Table 3-l) and ק (קשר) for that of *kéšer*$_\text{N}$ ‘relation’.

| | Lexeme | Morphological structure | Root type | Root | Morphophonological structure | Pattern-to-word alternation |
|---|---|---|---|---|---|---|
| a | limed ‘teach’ | CiCeC | r | l-m-d | \|ie\| | |
| b | nagar ‘carpenter’ | CaCaC | r | n-g-r | \|aa\| | |
| c | hirxiv ‘make wide’ | hiCCiC | r | r-x-v | hi\|0i\| | |
| d | hitraxev ‘become wide’ | hitCaCeC | r | r-x-v | hit\|ae\| | |
| e | nagariya ‘carpentry shop’ | CaCaC+iya | r | n-g-r | \|aa\|+iya | |
| f | yam ‘sea’ | raw | NA | NA | W | |
| g | spam ‘spam’ | borrowed | NA | NA | \|0a\| | |
| h | kéšer ‘relation’ | CéCeC | r | k-š-r | \|ée\| | |
| i | rac ‘run’ | CaCaC | r-C2=W | r-W-c | \|a\| | C2:0, V2:a~0 |
| j | pirsem ‘publish’ | CiCeC | r-4 | p-rs-m | \|ie\| | |
| k | siper ‘tell / cut hair’ | CiCeC | r-hom | s-p-r | \|ie\| | |
| l | kóšer ‘ability’ | CóCeC | r-hoph | k-š-r | \|óe\| | |

*Table 3. Word representation in Hebrewnette*

| | L1/L2 | L1/L2 Formal relation | L1/L2 Phonological variation | Root1/Root2 | Root1/Root2 Relation |
|---|---|---|---|---|---|
| a | kivec$_{\mathrm{V}}$/mexuvac$_{\mathrm{A}}$ | CiCeC/meCuCaC | C1:k~x | k-v-c / k-v-c | = |
| b | tadrix$_{\mathrm{N}}$/tidrex$_{\mathrm{V}}$ | taCCiC/CiCeC | – | d-r-x/t-dr-x | CCC / tCCC |
| c | spam$_{\mathrm{N}}$/hispim$_{\mathrm{V}}$ | CCaC/hiCCiC | – | NA/s-p-m | NA |

*Table 4. Phonological properties of the L1/L2 relations in Hebrewnette*

In addition to the annotations above, Hebrewnette describes **phonological variations** possibly involved by derivational relation (see Section 2.1.3), in line with the theoretical principles adopted in its conception, as mentioned in Section 2.3:

- between connected words - and more generally, the word patterns: in Table 4-a, there is a stop-fricative /k/~/x/ alternation affecting the first consonant of the pattern;
- between roots: a derivational relation may trigger the creation of a new branch in a derivational family, characterized by an additional consonant in the root shared by the members of this new sub-family. Such derived roots (e.g. t-dr-x, Table 4-b) are formed by prefixing the base root (e.g. d-r-x) with the consonant (e.g. t) of the prefix included in the pattern of the words with this base root (e.g. taCCiC); relations between roots are encoded according to the value of the new root element, when this is relevant (e.g. CCC/tCCC in Table 4-b, see Section 2.1.3);
- even when one of the two related words does not have a root (in Table 4-c), as in the rootless and patternless borrowed noun *spam*, cf. Table 3-f, it has an apparent phonological representation of the form CCaC, Table 4-c, consistent with the morpho-phonological representation |0a|, and containing the consonant cluster sp. This representation may be relevant and taken into consideration in the formation of a verb like *hispim*, that has the apparent root s-p-m[4], and looks like other native Hebrew verbs with root and pattern (e.g. *hidpis* 'print'). The root s-p-m and specifically the sp cluster are represented in the derived verb (Bat-El 1994, Bolozky 1978).

Regarding the **syntactic and semantic properties** of Hebrew words connected by paradigmatic relations (see Section 2), they are represented in Hebrewnette by several additional features (with respect to the ontological annotation already present in Démonette, see Section 3), where the main one describes verb argument structure. As

[4] We use the term "apparent root" to indicate that the base for word formation is either a raw native Hebrew word or a loan word (Section 2.2), which has no consonantal root. Since the formation of verbs based on such words must involve root and pattern morphology, the newly derived verbs seems to have a root. We thank an anonymous reviewer for the clarification.

| | L1 | | L2 | | L1 argument structure | L2 argument structure |
|---|---|---|---|---|---|---|
| a | lamad$_{\mathrm{V}}$ | 'learn' | nilmad$_{\mathrm{V}}$ | 'be learned' | XY | YX |
| b | limed$_{\mathrm{V}}$ | 'teach' | lamad$_{\mathrm{V}}$ | 'learn' | WXY | XY |
| c | lamad$_{\mathrm{V}}$ | 'learn' | hitlamed$_{\mathrm{V}}$ | 'train$_{intrans}$' | XY | X |
| d | limed$_{\mathrm{V}}$ | 'teach' | limud$_{\mathrm{N}}$ | 'teaching' | WXY | WXY |
| e | lamad$_{\mathrm{V}}$ | 'learn' | lamid$_{\mathrm{A}}$ | 'learnable' | XY | Y |

*Table 5. Morphologically predictable argument structure of predicates in Hebrewnette*

shown in Table 5, argument structure is encoded by means of the variables X, Y, W, which represent arguments of the predicate: in a family the same variable systematically corresponds to the same thematic role of the argument. For example, Table 5 displays an excerpt of the morphological network realized around the CaCaC transitive agentive verb *lamad* 'learn'.

- Table 5-a describes the relation between a transitive active (XY) structure and the corresponding passive one (YX), where X stands for the agent, and Y for the patient. Table 5-b describes the relation between a transitive causative verb (WXY) and its active transitive (XY) counterpart: W is an additional argument that causes the event represented by *lamad*.
- Table 5-c describes a relation between a transitive active agentive predicate (XY) and the corresponding intransitive verb (X).
- Moreover, since argument structure prediction goes beyond the verbal network in Semitic languages, the identity relation can also be defined between the structure of an active verb and that of its action noun (eg, Table 5-d, *limud*$_{\mathrm{N}}$, which inherits its argument structure from *limed*$_{\mathrm{V}}$), and the patient argument of a transitive verb can be passed to its related able-like adjective expressing potentiality (eg Table 5-e, *lamid*$_{\mathrm{A}}$, which inherits its external argument Y from the argument structure XY of *lamad*$_{\mathrm{V}}$).

In addition to the features and values which were presented, Hebrew morphology requires two distinct attributes to provide a precise description of the **orientation** of a relation: one for the morphological orientation, the other for the semantic one. This double organization, more complex than what is encoded in Démonette (see Table 1), is necessary for an accurate representation of form-meaning mismatches, as will be shown in Section 5.

## 5. Case studies

We now turn examine two case studies that deal with different theoretical aspects of Hebrew morphology that have been addressed in previous papers. We use these case studies to demonstrate how the properties of Hebrewnette allow us to provide empirical evidence in order to answer theoretical questions and shed light on the structural and semantic relations between words that are formed via non-concatenative morphology.

### 5.1. Faithfulness constraints and competing patterns

This case study examines doublet formation of Hebrew instrument nouns (hereafter INs). This is a case in which two INs that share the same meaning and consonantal root are constructed in two different patterns. Such variation is shown in Table 6.

| L1 | L2 | Gloss | Root | Pattern1 | Pattern2 |
|---|---|---|---|---|---|
| masnen | mesanen | 'filter'$_N$ | s-n-n | maCCeC | meCaCeC |
| maghec | megahec | 'iron'$_N$ | g-h-c | maCCeC | meCaCeC |
| magresa(t)[-kerax] | gores[-kerax] | '[ice-]crusher'$_N$ | g-r-s | maCCeCa | CoCeC |
| maxlec[-pkakim] | xolec[-pkakim] | '[cork-]screw'$_N$ | x-l-c | maCCeC | CoCeC |

*Table 6. Morphological variation of Instrument nouns (INs)*

As shown in Table 6, *masnen*, for example, is formed in maCCeC while *mesanen* is formed in meCaCeC, and both nouns share the consonants s-n-n and denote 'filter'. Nouns formed in the patterns of the Pattern1 column are typically considered the prescriptive forms, unlike nouns formed in the patterns of the Pattern2 column, which have become more frequent in current usage; speakers demonstrate a tendency to use the non-prescriptive forms to different degrees (Bolozky 1999, 2003). Regardless of the issue of the normative forms, both words share the same meaning and can be used in the same semantic-syntactic context. Laks (2015) shows that such doublet formation and lack thereof can be predicted based on morphological and semantic criteria. In this study, we address the morphological aspect of doublet formation and show that the properties of Hebrewnette allow us to predict which doublet member is preferred.

We begin with some background on Hebrew INs formation. There are two main groups of INs patterns. The participle patterns CoCeC, meCaCeC, maCCiC are ambiguous as they also denote the present tense of verbs, as illustrated in Table 7. The form *sorek*, for example, corresponds both to the noun 'scanner' and the present form of the verb *sarak* 'scan'.[5]

[5] The participle patterns can also denote agent nouns, e.g. *moxer* 'seller', related to the verb *maxar* 'sell', and also adjectives, e.g. *madhim* 'amazing', related to the verb *hidhim* 'amaze' (Section 2.1.2). Faust (2011) shows that agent nouns and INs are formed independently, i.e. without a corresponding verb, only in the CoCeC pattern. That is, other participle patterns do not host such independent nouns without a verbal alternate in the relevant pattern.

| | Verb Pattern | Example | | IN / Participle Pattern | Example | |
|---|---|---|---|---|---|---|
| a | CaCaC | sarak | 'scan' | CoCeC | sorek | 'scanner' |
| b | CiCeC | yibeš | 'dry' | meCaCeC | meyabeš | 'drier' |
| c | hiCCiC | hismix | 'thicken' (liquids) | maCCiC | masmix | 'thickener' |

*Table 7. INs formation in participle patterns*

Other patterns that host INs are not used as verbs and are not related to a specific verbal pattern (hereafter 'non-participle patterns'). Some of these patterns are presented in Table 8. It is important to note that this is not an exhaustive list, but it represents the common patterns in which INs are formed. Some of them, e.g. maCCeC, are more typical for INs than others, e.g. CaCaC, and none of them is exclusively used for INs formation (see Bolozky 1999, Schwarzwald 2002: and references therein). For example, the noun *mirpéset* 'balcony' is formed in the miCCeCet pattern, but denotes a location rather than an IN. This corresponds to a different ontological value in Hebrewnette.

| | Pattern | Example | |
|---|---|---|---|
| a | maCCeC | maxbet$_N$ | 'bat' |
| b | maCCeCa | maclema$_N$ | 'camera' |
| c | miCCéCet | miklédet$_N$ | 'keyboard' |
| d | maCCéCet | madpéset$_N$ | 'printer' |
| e | CaCaC | vasat$_N$ | 'regulator' |

*Table 8. INs formation in non-participle patterns*

INs doublets are formed in cases where an existing IN in a non-participle pattern takes an additional form in a participle pattern (Bolozky 2003, Laks 2015). This is shown in Table 9 where the form in (ii) is preferred over the form in (i) in both cases.

| | Verb | Verb Pattern | Instrument noun | | Nominal Pattern |
|---|---|---|---|---|---|
| a | sinen$_V$ 'filter' | CiCeC | (i) masnen<br>(ii) mesanen | 'filter' | maCCeC<br>maCaCeC |
| b | hidgiš$_V$ 'emphasize' | hiCCiC | (i) madgeš<br>(ii) madgiš | 'marker' | maCCeC<br>maCCiC |

*Table 9. Doublet formation of INs*

One of the doublets is preferred due to faithfulness to the base from which it is derived. In Table 9-a, the formation of *mesanen* is more faithful to *sinen* as it involves only prefixation and changing one vowel, while the prosodic structure remains intact. In contrast, the formation of *masnen* changes the prosodic structure of the base, as it creates the consonant cluster sn that does not exist in the verb *sinen*. In Table 9-b, the formation of both *madge*š and *madgi*š does not change the prosodic structure of the verb, but *madgi*š is more faithful because its second vowel /i/ is identical to the second vowel of the related verb *hidgi*š. The formation of both instrument nouns in (ii) involves fewer changes with respect to the verb, and as a result there is greater structural transparency between the verb and the instrument noun.

Hebrewnette provides the information required to compare concurrent INs according to their degree of faithfulness to their related verb, thanks to the value of the feature called Morphophonological structure we have introduced in Section 4, Table 3. As shown in Table 10, this feature allows computing the difference in the edit distance (known as Levenshtein distance) between the verb L1 and the 'regular' IN form L2 in Table 10-a,c on the one hand, and between L1 and the concurrent IN L2 in Table 10-b,d, on the other hand. The smaller the edit distance, the greater the faithfulness of L2 to L1, the more likely the L2 form.

We use a measure parametrized such that string modification is weighed according to the distance from the original syllabic and melodic structure. Therefore we decided that vowel substitution is twice as 'cheaper' (distance=1) as prefix insertion (or substitution or deletion) (distance=2). Moreover, it weights four times less than vowel insertion (or deletion) (distance=4), because the latter transformation involves "consonant (de)clusterization", that is, either breaking consonant clusters that exist in the base, or creating consonant clusters that are not part of the base.

The IN in Table 10-b is preferred over the one in Table 10-a because its edit distance from the verbal base is 3, while in Table 10-a it is 6. Similarly, the IN doublet member in Table 10-d is preferred over the one in Table 10-c because its edit distance from the verb is smaller.[6] As shown, the features encoded in Hebrewnette allow us to deduce predictions with respect to doublet formation and explain why one of the two doublets is preferred over the other.

Other theoretical hypotheses are empirically validated thanks to Hebrewnette annotations. This is what we show through a second case study.

### 5.2. Form/meaning mismatches

This second case study addresses transitivity alternations. Transitive-intransitive alternations within verbal systems have been the object of various studies including Alexiadou et al. (2006), Berman (1982, 1993), Borer (1991), Doron (2003), Haspel-

[6] Notice that both the Levenshtein measure used to predict the most likely IN competitor and the weight assigned to each criterion, in the two rightmost columns in Table 10 are external from the design of Hebrewnette and can be adjusted according to the need of the database users.

<table>
<tr><th rowspan="2"></th><th rowspan="2">L1</th><th rowspan="2">L2</th><th colspan="2">Morpho-phonological Structure</th><th rowspan="2">L1 / L2 changes</th><th rowspan="2">Edit distance</th></tr>
<tr><th>L1</th><th>L2</th></tr>
<tr><td>a</td><td>sinen$_V$</td><td>masnen$_N$</td><td>|ie|</td><td>ma|0e|</td><td>prefix insertion: ma-<br>vowel insertion: i</td><td>6</td></tr>
<tr><td>b</td><td>sinen$_V$</td><td>mesanen$_N$</td><td>|ie|</td><td>me|ae|</td><td>prefix insertion: me-<br>vowel change: i ↔ a</td><td>3</td></tr>
<tr><td>c</td><td>hidgiš$_V$</td><td>madgeš$_N$</td><td>hi|0i|</td><td>ma|0e|</td><td>prefix change: hi- ↔ ma-<br>vowel change: i ↔ e</td><td>3</td></tr>
<tr><td>d</td><td>hidgiš$_V$</td><td>madgiš$_N$</td><td>hi|0i|</td><td>ma|0i|</td><td>prefix change: hi- ↔ ma-</td><td>2</td></tr>
</table>

*Table 10. Predicting the outcome of INs competition in Hebrewnette*

math (1987), Horvath and Siloni (2008, 2010), Koontz-Garboden and Levin (2005), Levin and Rappaport Hovav (1995), Pylkkänen (2008), Reinhart (1996), Rappaport Hovav and Levin (2007, 2012), Williams (1981). It is commonly assumed that different thematic realizations of the same concept are not accidental and that there are some sort of derivational relations between verbs that participate in such alternations.

These alternations have been addressed by syntactic, semantic, and morphological theories, attempting to shed light on both the morphological and the syntactic and semantic-thematic characteristics of such derivations

Causative / inchoative alternations can involve apparent morpho-semantic mismatches, as discussed in Borer (1991), Doron (2003), Haspelmath (1987, 1993), Horvath and Siloni (2010), Rappaport Hovav and Levin (2012) among many others, where semantic and morphological directions seem to collide. In each pair of verbs in Table 11, the semantic relation is similar, where the transitive verbs denote causation of change in Y's mental state, and the intransitive verbs denote the change in the mental state that Y undergoes (see Table 5 for the way semantic roles are assigned to the symbols X, Y etc.). However, their structural (morphological) relations are different. In Table 11-a, the morphological relation is formally oriented from the transitive verb to the intransitive verb, as the former is formed in an affixless pattern (CiCeC), while the latter is formed in a pattern with a prefix (hitCaCeC). In contrast, in the relation of Table 11-b, the transitive verb, formed in a pattern with a prefix (hiCCiC), is formally more complex than the intransitive one formed in CaCaC.

To represent the form-meaning mismatch illustrated in Table 11, Hebrewnette encodes separately semantic and structural information about the direction of derivational relations. Based on the orientation attribute and its 'as2des' and 'des2as' values, presented for Démonette in Section 3 (see Table 1), orientation is duplicated into two properties: a formal and a semantic one. This organization enables an accurate representation of mismatches like with the two verb pairs of Table 11.

This is illustrated in Table 12. Since there are different approaches regarding semantic directionality for causative/inchoative alternations (see for example, the discussion in (Horvath and Siloni 2010)), we decided to encode semantic orientation as unspecified (`NA`) for both verb pairs, while the semantic difference between the transitive and the intransitive verb of each pair is given by the respective value of the argument-structure (in the `Argument Structure` columns) we introduced in Table 5.

In contrast, `Formal (or morphological) orientation` is determined by the presence of a prefix (and lack thereof) in one of the verbs in each pair. For the first verb pair, the `Formal Orientation` goes from `L1` to `L2` (`as2des`) because only `L2` (*hityaʔeš*) consists of a prefix (*hit-*), while the opposite orientation holds for the second verb pair, because only `L1` (*hitsis*) consists of a prefix (*hi-*). The value of the morphological orientation can be automatically computed using the same edit distance principles used for predicting the outcome of the competition between IN patterns (Section 5.1, Table 10), as shown in the last column of Table 12.

| Transitive V | | Pattern | Intransitive V | | Pattern |
|---|---|---|---|---|---|
| yiʔeš | ‘X make Y desperate’ | CiCeC | hityaʔeš | ‘Y become desperate’ | hitCaCeC |
| hitsis | ‘X make Y agitated’ | hiCCiC | tasas | ‘Y become agitated’ | CaCaC |

Table 11. Transitive/intransitive alternation

| L1 | L2 | Argument structure | | SO | Morpho-phonological structure | | FO | L1 / L2 Changes |
|---|---|---|---|---|---|---|---|---|
| | | L1 | L2 | | L1 | L2 | | |
| yiʔeš | hityaʔeš | XY | Y | NA | \|ie\| | hit\|ae\| | as2des | prefix insertion: hit-<br>vowel change: i ↔ a |
| hitsis | tasas | XY | Y | NA | hi\|0i\| | \|aa\| | des2as | prefix insertion: hi-<br>vowel insertion: a<br>vowel change: i ↔ a |

Table 12. Formal (FO) vs. semantic (SO) orientation in Hebrewnette

## 6. The **Hebrewnette** prototype

The above pilot study has resulted in the design of a prototypical database for Hebrew. The goal is to cover all the morphological phenomena that the complex lexicon of Hebrew may present, and more generally, all the paradigmatic aspects related to non-concatenative morphology. The collection of relations to be included in this prototypical version of the database combines two strategies.

- The first one is entirely manual; it consists in selecting a set of families according to the property(ies) that distinguish each of them, in order to test the expressive power of the notation system presented in Section 4.
- The second one is semi-automatic, and takes advantage of the partial predictability of verbal paradigms. It includes the automatic generation (followed by a manual post-editing) of families centered on pivot verbs instantiating the CiCeC pattern.

The first strategy follows "exemplar-based" principles. The word families that were selected intended to cover typical relations between Hebrew words. Therefore, the whole set of properties described in Sections 2, 4, and 5 correspond to at least one word-family included in the database, and include, among others:

- the type of root (Section 2.1.3 and Table 3),
- the mode of lexeme formation: relations between patterns (Section 2.1.1 and 2.1.2) and affixation (Section 2.2),
- the phonological alternation between a pattern and the form which realizes it (last column of Table 3),
- the pattern-to-pattern phonological variations (Table 4),
- the formation of subfamilies by root-to-root relations (Section 2.1.3 and Table 4),
- the form-meaning discrepancies (Section 5.2),
- the different cases of argument structure (Table 5).

This "family-centered" coverage describes the relations among a total of 245 lexemes belonging to 28 different families of different size (containing at least 4 members, as in the family of *yam*$_{\mathrm{N}}$ 'sea', and at most 28, as in the family of *kešer*$_{\mathrm{N}}$ 'relation').

The second strategy, of "pattern-centered" coverage, is based on the regularities observed empirically in families based on CiCeC verbs. CiCeC has been selected because it is the most productive pattern for newly coined verbs and subsequent families, with a relatively higher predictability. In contrast, other verbal patterns (CaCaC, niCCaC) are either unproductive (few new families come from verbs in the CaCaC pattern) or unpredictable: apart from the active-passive relation, it is more difficult to predict the content and size for families of verbs in the other patterns, that is, hiCCiC or hitCaCeC. Regularities regarding CiCeC verbs have been encoded in order to semi-automatically generate and annotate derivational families based on an initial list of 10 CiCeC verbs. Relying on the fundamentally paradigmatic nature of the Hebrew verbal lexicon the following predictions have been implemented:

- CiCeC verbs are likely to realize active, transitive, dynamic predicates, e.g. *xibek* 'hug$_V$', *kivec* 'shrink$_V$', *nihel* 'manage$_V$';
- they are related to a CiCuC action noun (*xibuk* 'hug$_N$', *nihul* 'management$_N$'), a resultative adjective in the meCuCaC participle pattern (*menohal*[7] 'managed$_A$'). CiCeC is also derivationally related to the meCaCeC participle pattern that can surface as an agent noun (*menahel* 'manager$_N$') an instrument noun, or an adjective (Section 2.1.2);
- when it is attested, their hitCaCeC related verb is intransitive, typically inchoative (*hitkavec* 'become shrunk$_V$'), reflexive (*hitraxec* 'wash oneself$_V$') or reciprocal (*hitxabek*, 'hug each other$_V$').

From these 10 CiCeC verbs, the program produced 70 new annotated lexemes (after manual verification, 20 of them are discarded): each CiCeC verb is the source of a family of 6 members on average. As each member in a family is linked to all the others, this amounts to supplementing the 245 initial word pairs with 90 new fully documented entries.

## 7. Conclusions

This paper presented the main principles of the architecture of Hebrewnette, a derivational database for Hebrew, and its properties. We accounted for the adaptations that were made on the Démonette database, which was originally designed for Romance Morphology. Focus was on non-concatenative formation, which is highly typical of Hebrew and Semitic languages in general. We outlined the way words were coded with respect to their root and pattern. Taking a word-based approach for word formation, Hebrewnette is based on coding relations between words, and specifically for Hebrew, relations between roots and patterns. It is based on separate descriptions of semantic and structural relations so that each type of relation can be examined according to different criteria, e.g. direction of derivation and morphophonological alternations (if any). The features and feature values in the Hebrewnette database intertwine with the content of Démonette, to account for the properties of non-concatenative morphology. However, these additions do not compromise the architecture of Démonette; the global structures of the two databases are superimposable, which allows us to envisage a total interoperability between the two systems (and more generally between the morphologies of Romance languages and Semitic languages). We examined two cases that demonstrate how generalizations about the nature of Hebrew morphology can be captured based on the properties of Hebrewnette. The case of doublet formations of instrument nouns demonstrates the importance of the Hebrewnette representation of structural relations between words and how such representations can provide predictions with respect to the likelihood of

[7] The /u/ to /o/ variation between the pattern meCuCaC and the word *menohal* is due to the fact that the second consonant of the root /h/ is a glottal.

doublets to be derived. The case of form/meaning mismatches in transitivity alternations sheds light on the importance of distinguishing between semantic and structural descriptions of relations between words and the relevant implementation of describing relations between words that are formed in patterns.

## Acknowledgements

This research has been realized as part of the project Demonext[8], supported by the Agence Nationale de la Recherche, grant number: ANR-17-CE23-0005. It has been also been supported by the Chateaubriand Fellowship Program of the French Embassy in Israel.

## Bibliography

Alexiadou, Artemis. *Functional structure in nominals: nominalization, and ergativity*. John Benjamins, Amsterdam, 2001. doi: 10.1075/la.42.

Alexiadou, Artemis, Elena Anagnostopoulou, and Florian Schäfer. The properties of anticausatives crosslinguistically. In Frascarelli, Mara, editor, *Phases of Interpretation*, pages 187–212. Mouton, Berlin, 2006. doi: 10.1515/9783110197723.4.187.

Anderson, Stephen R. *A-Morphous Morphology*. Cambridge University Press, Cambridge, UK, 1992. doi: 10.1017/CBO9780511586262.

Arad, Maya. *Roots and Patterns: Hebrew morpho-syntax*. Springer, Dordrecht, 2005.

Aronoff, Mark. *Word Formation in Generative Grammar*. Linguistic Inquiry Monographs. MIT Press, Cambridge, MA, 1976.

Aronoff, Mark. *Morphology by Itself. Stems and Inflectional Classes*. MIT Press, Cambridge, MA, 1994.

Aronoff, Mark. In the Beginning was the Word. *Language*, 83:803–830, 2007. doi: 10.1353/lan.2008.0042.

Bat-El, Outi. *Extraction in Modern Hebrew morphology*. Phd thesis, UCLA, California, 1986.

Bat-El, Outi. Stem modification and cluster transfer in Modern Hebrew. *Natural Language and Linguistic Theory*, 12:572–596, 1994. doi: 10.1007/BF00992928.

Bat-El, Outi. Semitic verb structure within a universal perspective. In Shimron, Joseph, editor, *Languages processing and acquisition in languages of semitic, root-based, morphology*, pages 29–59. Benjamins, Amsterdam, 2002. doi: 10.1075/lald.28.02bat.

Bat-El, Outi. Prosodic alternations in Modern Hebrew segolates. In Muchnik, Malka and Zvi Sadan, editors, *Studies on Modern Hebrew and Jewish Languages*, pages 116–129. Carmel Press, Jerusalem, 2012.

Bat-El, Outi. Word-based items-and processes (WoBIP): Evidence from Hebrew morphology. In Bowern, Claire, Laurence Horn, and Raffaella Zanuttini, editors, *On Looking into Words (and beyond)*, pages 115–135. Language Science Press, Berlin, 2017.

[8] https://www.demonext.xyz/

Bat-El, Outi. Templatic morphology (clippings, root-and-pattern). In *Oxfors Research Encyclopedia of Linguistics*. Oxford University Press, Oxford, 2019.

Batsuren, Khuyagbaatar, Gábor Bella, and Fausto Giunchiglia. MorphyNet: a Large Multilingual Database of Derivational and Inflectional Morphology. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 39–48, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.sigmorphon-1.5. URL `https://aclanthology.org/2021.sigmorphon-1.5`.

Becker, Thomas. Back-formation, cross-formation, and 'bracketing paradoxes' in paradigmatic morphology. *Yearbook of Morphology 1992*, pages 1–25, 1993. doi: 10.1007/978-94-017-3712-8_1.

Benmamoun, Elabbas. The role of the imperfective template in Arabic morphology. In Shimron, Joseph, editor, *Language Processing and Acquisition in languages of Semitic, root-Based, morphology*, pages 99–114. John Benjamins, Amsterdam, 2003. doi: 10.1075/lald.28.05ben.

Berman, Ruth. Modern Hebrew structure. Report, University Publishing Projects, 1978.

Berman, Ruth. Verb-pattern alternation: the interface of morphology, syntax, and semantics in Hebrew child language. *Journal of Child Language*, 9:169–191, 1982. doi: 10.1017/S030500090000369X.

Berman, Ruth. Marking of verb transitivity by Hebrew-speaking children. *Journal of Child Language*, 20:641–669, 1993. doi: 10.1017/S0305000900008527.

Berman, Ruth and Batia Seroussi. Derived nouns in Hebrew: Structure, meaning, and psycholinguistic perspectives. *Italian Journal of Linguistics*, 23(1):105–125, 2011.

Bolozky, Shmuel. Word formation strategies in MH verb system: denominative verbs. *Afroasiatic Linguistics*, 5:1–26, 1978.

Bolozky, Shmuel. The segolates: Linear or discontinuous derivation? In Schwarzwald, Ora R. and Izchak M. Schlesinger, editors, *Hadassah Kantor Jubilee Book* (*in Hebrew*). Bar-Ilan University, Ramat Gan, 1995.

Bolozky, Shmuel. *Measuring productivity in word formation: the case of Israeli Hebrew*. Brill, Leiden, 1999. doi: 10.1163/9789004348431.

Bolozky, Shmuel. Phonological and morphological variations in spoken Hebrew. In Hary, Benjamin H., editor, *Corpus Linguistics and Modern Hebrew*, pages 119–156. Rosenberg School of Jewish Studies, Tel Aviv, 2003.

Bolozky, Shmuel. More on linear vs. discontinuous derivation in Israeli Hebrew morphology. In Muchnik, Malka and Zvi Sadan, editors, *Studies in Modern Hebrew and Jewish*, pages 50–59. Carmel, Jerusalem, 2012.

Bonami, Olivier and Jana Strnadová. Paradigm structure and predictability in derivational morphology. *Morphology*, pages 1–31, 2018. doi: 10.1007/s11525-018-9322-6.

Borer, Hagit. The causative-inchoative alternation: a case study in parallel morphology. *The Linguistic Review*, 8:119–158, 1991. doi: 10.1515/tlir.1991.8.2-4.119.

Borer, Hagit. Derived Nominals and the Domain of Content. *Lingua*, 141:71–96, 2014. doi: 10.1016/j.lingua.2013.10.007.

Comrie, Bernard and Sandra A. Thompson. Lexical Nominalization. In Shopen, Thimothy, editor, *Language Typology and Syntactic Description*, volume III, pages 334–381. Cambridge University Press, Cambridge, 2007. doi: 10.1017/CBO9780511618437.006.

Corbett, Greville G. Canonical derivational morphology. *Word Structure*, 3(2):141–155, 2010. doi: 10.3366/word.2010.0002.

Davis, Stuart and Bushra Adnan Zawaydeh. Arabic hypocoristics and the status of the consonantal Root. *Linguistic Inquiry*, 32(3):512–520, 2001. doi: 10.1162/002438901750372540.

Daya, Ezra, Dan Roth, and Shuly Wintner. Identifying Semitic Roots: Machine Learning with Linguistic Constraints. *Computational Linguistics*, 34(3):429–448, 2008. doi: 10.1162/coli.2008.07-002-R1-06-30.

Doron, Edit. Agency and voice: The semantics of the Semitic templates. *Natural Language Semantics*, 11:1–67, 2003.

Farwaneh, Samira. Well-formed associations in Arabic: Rule or condition. In Eid, Mushira and John McCarthy, editors, *Perspectives on Arabic Linguistics*, volume II, pages 120–142. John Benjamins, Amsterdam/ Philadelphia, 1990. doi: 10.1075/cilt.72.08far.

Faust, Noam. *La Morpho-Syntaxe Nominale de l'hebreu moderne du point de vue de la forme phonologique*. Thèse de doctorat, Université Paris-Diderot, 2011.

Faust, Noam. A novel, combined approach to Semitic word-formation. *Journal of Semitic Studies*, LX(2):287–316, 2015. doi: 10.1093/jss/fgv001.

Faust, Noam. New reasons to root for the Semitic root from Mehri and Neo-Aramaic. *The Linguistic Review*, 36(3):575–599, 2019. doi: 10.1515/tlr-2019-2030.

Faust, Noam and Yarar Hever. Empirical and Theoretical Arguments in Favor of the Discontinuous Root in Semitic Languages. *Brill's Journal of Afroasiatic Languages and Linguistics*, 2: 80–118, 2010. doi: 10.1163/187666310X12688137960704.

Goldenberg, Gideon. Principles of Semitic word-structure. In Goldenberg, Gideon and Schlomo Raz, editors, *Semitic and Cushitic Studies*, pages 10–45. Harrassowitz Verlag, Wiesbaden, 1994.

Grimshaw, Jane. *Argument Structure*. MIT Press, Cambridge (MA) / London, 1990.

Hammond, Michael. Templatic transfer in Arabic broken plurals. *Natural Language and Linguistic Theory*, 6:247–270, 1988. doi: 10.1007/BF00134231.

Haspelmath, Martin. Transitivity alternations of the anticausative type. In *Institut für Linguistik (Köln). Abteilung Allgemeine Sprachwissenschaft: Arbeitspapier ; N.F., Nr. 5*. 1987.

Haspelmath, Martin. *More on the typology of inchoative / causative verb alternations*, pages 87–111. John Benjamins, Amsterdam/Philadelphia, 1993. doi: 10.1075/slcs.23.05has.

Hathout, Nabil. Une approche topologique de la construction des mots : propositions théoriques et application à la préfixation en *anti-*. In *Des unités morphologiques au lexique*, pages 251–318. Hermès Science-Lavoisier, Paris, 2011.

Hathout, Nabil and Fiammetta Namer. Discrepancy between form and meaning in Word Formation: the case of over- and under-marking in French. In Rainer, Franz, Wolfgang U. Dressler, Francesco Gardani, and Hans Christian Luschützky, editors, *Morphology and meaning*, pages 177–190. John Benjamins, Amsterdam, 2014. doi: 10.1075/cilt.327.12hat.

Hathout, Nabil and Fiammetta Namer. Giving Lexical Resources a Second Life: Démonette, a Multi-sourced Morpho-semantic Network for French. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (*LREC 2016*), pages 1084–1091, 2016.

Hathout, Nabil and Fiammetta Namer. ParaDis: a family and paradigm model. *Morphology*, 2022. doi: 10.1007/s11525-021-09390-w. URL `https://doi.org/10.1007/s11525-021-09390-w`.

Hazout, Ilan. Action nominalizations and the lexicalist hypothesis. *Natural Language and Linguistic Theory*, 13:355–404, 1995. doi: 10.1007/BF00992736.

Heath, Jeffrey. *Ablaut and ambiguity: Phonology of a Moroccan Arabic dialect*. State University of New York Press, Albany, 1987.

Hoberman, Robert D. Local spreading. *Journal of Afroasiatic Linguistics*, 3:226–254, 1992.

Hockett, Charles Francis. Two models of linguistc descriptions. *Words*, 10:210–234, 1954.

Horvath, Julia and Tal Siloni. *Active lexicon: Adjectival and verbal passives*, pages 105–134. John Benjamins, Amsterdam/ Philadelphia, 2008. doi: 10.1075/la.134.05act.

Horvath, Julia and Tal Siloni. *Lexicon versus syntax: Evidence from Morphological Causatives*, pages 153–176. Oxford University Press, Oxford, 2010.

Iacobini, Claudio. Parasynthesis in Morphology. In *Oxford Research Encyclopedia of Linguistics*. Oxford University Press, 2020. doi: 10.1093/acrefore/9780199384655.013.509.

Idrissi, Ali and Eva Kehayia. Morphological units in the Arabic mental lexicon: Evidence from an individual with deep dyslexia. *Brain and Language*, 90:183–197, 2004. doi: 10.1016/S0093-934X(03)00431-0.

Itai, Alon and Shuly Wintner. Language resources for Hebrew. *Language Resources and Evaluation*, 42:75–98, 2008. doi: 10.1007/s10579-007-9050-8.

Jackendoff, Ray. Morphological and semantic regularities in the lexicon. *Language*, 51(3):639–671, 1975. doi: 10.2307/412891.

Kastner, Itamar. Templatic morphology as an emergent property: Roots and functional heads in Hebrew. *Natural Language and Linguistic Theory*, 37(2):571–619, 2019. doi: 10.1007/s11049-018-9419-y.

Kastner, Itamar. *Voice at the interfaces: The syntax, semantics and morphology of the Hebrew verb*. Language Science Press, Berlin, 2020. doi: 10.5281/zenodo.3865067.

Kihm, Alain. Plural formation in Nubi and Arabic: A comparative study and a word-based approach. *Brill's Journal of Afroasiatic Languages and Linguistics*, 3:1–21, 2011. doi: 10.1163/187666311X562431.

Kirov, Christo, Ryan Cotterell, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sebastian Mielke, Arya D. McCarthy, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. UniMorph 2.0: Universal Morphology. In Calzolari, Nicoletta, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation* (*LREC 2018*), Miyazaki, Japan, May 7-12, 2018 2018. European Language Resources Association (ELRA). ISBN 979-10-95546-00-9.

Klimek, Bettina, Natanel Arndt, Sebastian Krause, and Timotheus Arndt. Creating Linked Data Morphological Language Resources with MMoOn. The Hebrew Morpheme Inventory. In Calzolari, Nicoletta, Khalid Choukri, Thierry Declerck, Sara Goggi, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Helene Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 892–899. European Language Resources Association (ELRA), 2016.

Koontz-Garboden, Andrew and Beth Levin. *The morphological typology of change of state event encoding*, pages 185–194. Università degli Studi di Bologna, Bologna, 2005.

Kyjánek, Lukáš. Morphological Resources of Derivational Word-Formation Relations. Technical Report 61, ÚFAL - Charles University, Prague, 2018.

Kyjánek, Lukáš, Zdeněk Žabokrtský, Magda Ševčíková, and Jonáš Vidra. Universal Derivations 1.0, A Growing Collection of Harmonised Word-Formation Resources. *The Prague Bulletin of Mathematical Linguistics*, 115:5–30, October 2020. ISSN 0032-6585. doi: 10.14712/00326585.003. URL `https://ufal.mff.cuni.cz/pbml/115/art-kyjanek-et-al.pdf`.

Laks, Lior. Variation and change in instrument noun formation in Hebrew and its relation to the verbal system. *Word Structure*, 8(1):1–28, 2015. doi: 10.3366/word.2015.0071.

Laks, Lior and Fiammetta Namer. Designing a derivational resource for non-concatenative Morphology: the Hebrewnette database. In Namer, Fiammetta, Nabil Hathout, Stéphanie Lignon, Zdeněk Žabokrtský, and Magda Ševčíková, editors, *DeriMo2021 - Third International Workshop on Resources and Tools for Derivational Morphology.*, pages 95–105. ATILF, 2020.

Levin, Beth and Malka Rappaport Hovav. *Unaccusativity*. MIT Press, Cambridge, MA, 1995.

Levin, Beth and Malka Rappaport Hovav. *Argument Realization*. Cambridge University Press, Cambridge, 2005. doi: 10.1017/CBO9780511610479.

Matthews, Peter H. *Inflectional Morphology*. Cambridge University Press, Cambridge, 1972.

Matthews, Peter Hugoe. *Morphology: an introduction to the theory of word-structure*. Cambridge University Press, London, 1974.

McCarthy, Arya D., Miikka Silfverberg, Ryan Cotterell, Mans Hulden, and David Yarowsky. Marrying Universal Dependencies and Universal Morphology. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 91–101, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6011. URL `https://aclanthology.org/W18-6011`.

McCarthy, John. A prosodic theory of nonconcatenative morphology. *Linguistic Inquiry*, 12: 373–418, 1981.

McCarthy, John and Alan Prince. *Prosodic morphology*. Thesis, 1986.

McCarthy, John and Alan Prince. Foot and word in Prosodic Morphology: The Arabic broken plural. *Natural Language and Linguistic Theory*, 8:209–283, 1990. doi: 10.1007/BF00208524.

Melloni, Chiara. *Event and Result Nominals: A Morpho-semantic Approach*. Peter Lang, Bern, 11 2011. ISBN 978-3-0343-0658-4.

Namer, Fiammetta and Nabil Hathout. ParaDis and Démonette –From Theory to Resources for Derivational Paradigms. *The Prague Bulletin of Mathematical Linguistics*, 114:5–33, 2020. doi: 10.14712/00326585.001.

Neme, Alexis Amid. A lexicon of Arabic verbs constructed on the basis of Semitic taxonomy and using finite-state transducers. In Sagot, Benoît, editor, *WoLeR 2011*, pages 79–86, 2011. URL halshs-01186723.

Nevins, Andrew. Overwriting does not optimize in non-concatenative word formation. *Linguistic Inquiry*, 36(2):275–287, 2005.

Nir, Bracha, Brian MacWhinney, and Shuly Wintner. The Hebrew CHILDES corpus: transcription and morphological analysis. *Language Resources and Evaluation*, 47(4):973–1005, 2013.

Ornan, Uzzi. *How is the Hebrew word formed?*, pages 13–42. Magnes, Jerusalem, 1983.

Pylkkänen, Liina. *Introducing arguments*. MIT Press, Cambridge, 2008. doi: 10.7551/mitpress/9780262162548.001.0001.

Rappaport Hovav, Malka and Beth Levin. *Deconstructing thematic hierarchies*, pages 385–402. CSLI Publications, Stanford, 2007.

Rappaport Hovav, Malka and Beth Levin. *Lexicon uniformity and the causative alternation*, pages 150–176. University Press, Oxford, 2012.

Rasin, Ezer, Omer Preminger, and David Pesetsky. A re-evaluation of Arad's Argument for Roots. In *Proceedings of the 39th West Coast Conference on Formal Linguistics*, 2021. URL https://ling.auf.net/lingbuzz/006077.

Ratcliffe, Robert R. *Prosodic templates in a word-based morphological analysis of Arabic*, pages 147–171. John Benjamins, Amsterdam/Philadelphia, 1997. doi: 10.1075/cilt.153.10rat.

Ravid, Dorit. Internal structure constraints on new-word formation devices in Modern Hebrew. *Folia Linguistica*, 24:289–347, 1990. doi: 10.1515/flin.1990.24.3-4.289.

Ravid, Dorit and Avraham Avidor. Acquisition of derived nominals in Hebrew: developmental and linguistic principles. *Journal of Child Language*, 25:229–266, 1998. doi: 10.1017/S0305000998003419.

Ravid, Dorit, Orit Ashkenazi, Ronit Levie, Galit Ben Zadok, Tehila Grunwald, Ron Bratslavsky, and Steven Gillis. Foundation of the Early Root Category: Analyses of Linguistic Input to Hebrew-Speaking Children. In Berman, Ruth, editor, *Acquisition and Development of Hebrew: From Infancy to Adolescence*, pages 95–134. John Benjamins, Amsterdam, 2016.

Reinhart, Tanya. Syntactic effects of lexical operations: Reflexives and unaccusatives. Report, University of Utrecht, 1996.

Rose, Sharon. Triple take: Tigre and the case of internal reduplication. In Hayward, Hay, Jamal Ouhalla, and Denise Perett, editors, *Studies in Afroasiatic grammar*. John Benjamins, Amsterdam, 1998.

Schwarzwald, Ora R. *Grammar and Reality in the Hebrew Verb*. Bar Ilan University Press, Ramat Gan, 1981.

Schwarzwald, Ora R. *Studies in Hebrew Morphology (in Hebrew)*. The Open University, Tel Aviv, 2002.

Singh, Nimesh and Nizar Habash. Hebrew Morphological Preprocessing for Statistical Machine Translation. In Cettolo, Mauro, Marcello Federico, Lucia Specia, and Andy Way, editors, *Proceedings of the 16th EAMT Conference*, pages 43–50. European Association for Machine Translation, 2012.

Spencer, Andrew. *Lexical relatedness*. Oxford University Press, Oxford, 2013. doi: 10.1093/acprof:oso/9780199679928.003.0003.

Steriade, Donca. Reduplication and transfer in Sanskrit and elsewhere. *Phonology*, 5(1):73–155, 1988.

Tribout, Delphine. *Les conversions de nom à verbe et de verbe à nom en français*. Thèse de doctorat, Université Paris 7, 2010.

Ussishkin, Adam. A fixed prosodic theory of nonconcatenative templatic morphology. *Natural Language and Linguistic Theory*, 23:169–218, 2005.

Williams, Edwin. Argument structure and morphology. *The Linguistic Review*, 1:81–114, 1981. doi: 10.1515/tlir.1981.1.1.81.

Wintner, Shuly. Hebrew computational linguistics: Past and future. *Artificial Intelligence Review*, 21(2):113–138, 2004. doi: 10.1023/B:AIRE.0000020865.73561.bc.

Yip, Moira. Template morphology and the direction of association. *Natural Language and Linguistic Theory*, 6:551–577, 1988. doi: 10.1007/BF00134493.

**Address for correspondence:**
Fiammetta Namer
fiammetta.namer@univ-lorraine.fr
Université de Lorraine, CLSH
UFR SHS, dépt SDL
23 bd Albert 1er, BP 60446
54001 NANCY CEDEX, France

PBML

The Prague Bulletin of Mathematical Linguistics

NUMBER 118 APRIL 2022 55–73

# Word Formation Analyzer for Czech: Automatic Parent Retrieval and Classification of Word Formation Processes

Emil Svoboda, Magda Ševčíková

Charles University, Faculty of Mathematics and Physics

**Abstract**

We present a deep-learning tool called *Word Formation Analyzer for Czech*, which, given an input lexeme, automatically retrieves the lemma or lemmas from which the input lexeme was formed. We call this task parent retrieval. Furthermore, based on the number of words in the output sequence and its comparison to the input, the input word is classified into one of three categories: *compound*, *derivative* or *unmotivated*. We call this task word formation classification. In the task of parent retrieval, *Word Formation Analyzer for Czech* achieved an accuracy of 71%. In word formation classification, the tool achieved an accuracy of 87%.

## 1. Introduction

A native speaker of Czech, when given a word, generally finds it easy to determine which Czech word or words it comes from, or if any such ancestor word exists. In contrast, there is no trivial automatic procedure that can do the same.

Research on this topic has so far been mostly limited to creating static data resources, similar in principle to dictionaries, capturing Czech words with links to their respective ancestors. The problem is that speakers and writers coin new words to suit their communicative needs; this implies that no static data resource can capture the entirety of Czech word formation at any given point in time. This creates the need for a procedural tool capable of handling any word, regardless if it is a long-established word or a new coinage.

Corresponding author: svoboda@ufal.mff.cuni.cz

Cite as: Emil Svoboda, Magda Ševčíková. Word Formation Analyzer for Czech: Automatic Parent Retrieval and Classification of Word Formation Processes. The Prague Bulletin of Mathematical Linguistics No. 118, 2022, pp. 55–73. doi: 10.14712/00326585.019.

In this paper, we present *Word Formation Analyzer for Czech* (*WFA.ces*), a tool based around an ensemble of three sequence-to-sequence deep-learning models. The tool takes as its input a string of characters assumed to be a Czech lexeme in its dictionary form (lemma), and returns a predicted sequence of one or more words the input lexeme was motivated by. Since the tool receives nothing but an isolated string as its input, the procedure is entirely based on the written form of the input. *WFA.ces* can perform two tasks:

1. *Parent retrieval*
   *WFA.ces* predicts which word or words the input lemma is motivated by. It does this by generating a list of candidate sequences of parent words, and returning the best sequence based on a particular reranking procedure of the user's choice. This task is similar to that of *stemming*, but with a stronger focus on linguistic adequacy.
2. *Word formation classification*
   *WFA.ces* classifies the input lexeme into one of the classes *compound*, *derivative*, or *unmotivated*. It returns the class *compound* if there are two or more words in the output (*hlavonožec* ('cephalopod') ← *hlava* ('head') + *noha* ('leg')); the class *derivative* if there is one word AND it differs from the input (*hlavička* ('little_head') ← *hlava* ('head')); and finally, if there is one word AND it is identical to the input, it returns the class *unmotivated* (*hlava* ('head') ← *hlava* ('head')).

For the purposes of our solution, we consider products of conversion in Czech to be derivatives. The reasoning behind this will be expanded upon further in Section 2. Similarly, we consider loanwords to be unmotivated, even in cases where they are clearly motivated in their source languages (cf. *downsizing* from English, *majstrštyk* from German, or *špageta* from Italian). Due to the retroactive nature of parent retrieval and word formation classification, all examples of word formation from here on out will be structured with the product word on the left side, followed by a leftwise arrow, with the parent(s) on the right side; cf. (1).

(1) ***product*** ← ***parent***$_1$ ***parent***$_2$
translation.POS translation.POS translation.POS

We begin this paper by briefly outlining the challenges of Czech word formation, especially derivation and compounding in Section 2. Section 3 relays the handling of these issues in natural language processing (NLP), and describes in brief the *Czech Compound Splitter* tool, which is the predecessor of *WFA.ces*. Next, in Section 4, we outline the various formal difficulties that Czech word formation presents, the data that was used to train the deep-learning model ensemble, and the evaluation metrics used to measure its performance. Section 5 presents the way the underlying ensemble was trained, how it functions, and how it ended up performing, including error analysis. Section 6 compares *WFA.ces* to its predecessor and outlines future research. Finally, Section 7 contains the summary of this paper.

## 2. Word Formation in Czech

The foundations of theoretical approaches toward word formation in Czech have been laid by Dokulil (1962) and, since then, broadly accepted and applied to Czech and other, particularly (but not exclusively) Slavic languages; cf. all reference grammars of Czech, including the representative volume by Dokulil et al. (1986) and the latest grammars by Štícha et al. (2013) or Štícha et al. (2018).

### 2.1. Derivation

The basic concept of derivation as a process of the formation of new words by adding derivational affixes to already-existing lexemes or roots is in Czech fundamentally complicated by allomorphy, homonymy, and other issues, which are difficult to model computationally. For instance, two different variants of the prefix (*vy-*, *vý-*) and three different allomorphs of the same root occur in the adjective *vybraný* 'chosen' (root *-br-*), in the noun *výběr* 'choice' (*-běr-*), and in the noun *výbor* 'committee' (*-bor-*), even if they are all motivated by the verb *vybrat* 'to choose', which is, in turn, based on *brát* 'to take'. Although verb prefixation is among the less irregular processes with a minimum of formal changes, problems can also be found here. An example is the verb *obléci* 'to dress', whose simple deprefixation yields a string that does not match any existing verb (cf. **bléci* or **léci* as both the prefix *o-* and *ob-* exist in Czech). This verb is to be traced back to the verb *vléci* 'to pull', in which the initial consonant is dropped when combined with the prefix *ob-* (because of the pronunciation: *ob+vléci*; cf. (2)), but remains in place with other prefixes ((3) and (4)).

(2) ***obléci*** ← ***vléci***
dress.VERB pull.VERB

(3) ***navléci*** ← ***vléci***
pull on.VERB pull.VERB

(4) ***svléci*** ← ***vléci***
take off (clothes).VERB pull.VERB

Circumfixation, understood as prefixation and suffixation in a single step, also occurs in Czech (5). This presents difficulty for automatic solutions, because in affixation mostly a single affix is added in each step. However, if derivation of the adjective *přidrzlý* is interpreted as a sequence of derivations (cf. (6) or alternatively (7)), the product of the middle step is unattested, and therefore an incorrect retrieval. An algorithmic solution, nevertheless, has no way of inferring attestability without consulting a corpus. The implementation of a corpus lookup can mitigate this particular problem, but may introduce other issues, as demonstrated in Section 5.2.

(5) ***přidrzlý*** ← ***drzý***
a bit cheeky.ADJ cheeky.ADJ

(6) ***přidrzlý*** ← ****přidrznout*** ← ***drzý***
a bit cheeky.ADJ become cheeky.VERB cheeky.ADJ

(7) ***přidrzlý*** ← ****drzlý*** ← ***drzý***
a bit cheeky.ADJ having become cheeky.ADJ cheeky.ADJ

In Czech, conversion is formally very similar to derivation in many cases.[1] The two processes differ solely by the type of affixes used. While derivational affixes are added in derivation, conversion is assumed to be the sole addition of inflectional morphemes without adding derivational affixes. For example, the adjective in (8) is considered to be converted from the noun, despite the fact that we see a total of *two* formal changes to the parent word. First, it is vowel deletion, which can also be see as an alternation (∅ ← /e/), which is common across all of Czech word formation (cf. (17) and (19) for examples in compounding), and the addition of the adjectival ending *-í*.

(8) ***psí*** ← ***pes***
dog.ADJ dog.NOUN

However, this relatively clear distinction is very difficult for automatic analysis. An example pipeline capable of doing so would require the following:

1. reliable morphological segmentation so as to isolate the morphemes of both the input and output words;
2. reliable morpheme alignment of the input and output morphemes onto each other in order to determine which morphemes, if any, were added;
3. reliable classification of the added morphemes as either *derivational* or *inflectional*.

Additionally, the mere changing of the POS and/or the inflectional pattern of a given word without any formal changes is in the Czech linguistic tradition also considered to be conversion. This is more akin to what is considered conversion in English (cf. (9)). Such a word formation procedure cannot, however, in principle be handled by a tool like *WFA.ces* because it accepts isolated lemmas represented by a string only. From the sole lemma, it is undecidable whether we mean *raněný* 'wounded' the noun, whose parent is *raněný* 'wounded' the adjective (10), or if we mean *raněný* the adjective, whose parent is the verb *ranit* 'to wound', as these need syntactic context to be disambiguated. Therefore, when the word *raněný* is passed into *WFA.ces*, the tool is expected to return *ranit*.

[1] Most cases of conversion in Czech, as in other inflecting languages, do not conform to the central type of conversion, which is characterized by the formal identity of the input and output lexeme (word-based conversion), but rather belong to the non-central type of conversion, where the input and output share the root but may differ in inflectional markers (root-based conversion; Valera and Ruz 2021).

(9) ***raněný*** ← ***raněný***
wounded.NOUN wounded.ADJ

(10) ***raněný*** ← ***ranit***
**wounded.ADJ** wound.VERB

For the reasons stated in the previous paragraphs, we have decided to consider conversion as derivation and to label it as such. From a theoretical perspective, this decision can be viewed as the interpretation of conversion as derivation by zero affix.

### 2.2. Compounding

Bozděchová (1997) distinguishes two types of compounding in Czech, depending on whether the words entering the composition are formally modified or not. *Compounding proper*, which requires morphological adjustment of the input words, and *compounding improper*, which is the result of simple concatenation of a syntactic phrase with no morphological adjustments. In addition, Bozděchová puts forth a multi-level classification, starting from the part-of-speech category of the output compound and then proceeding to semantic criteria (considering the meanings of the input items, of the output compounds and the relationship between the output and the inputs).

In a recent paper on compounding in West Slavic languages, Ološtiak and Vojteková (2021) focus on compounds partially or fully motivated by elements of Greek-Latin origin (from here: *neoclassical compounds*). Four types of word formation formants are distinguished, namely bases, baseoids, affixoids, and affixes. Bases are items that can appear freely and are lexically specific (*terapie* 'therapy', like in *ergoterapie* 'occupational therapy'); baseoids are items that do not appear freely, but are lexically specific regardless (*ergo-*, in *ergoterapie* 'occupational therapy'); affixoids are non-independent items which have gradually lost their lexical specificity (*-náct* like in *třináct* 'thirteen' – originally from *na deset* 'to_ten'); and affixes are items which carry lexically non-specific meaning, referencing a group of referents within a given part of speech, like "object", "place", "tool", "agent" for nouns (*-ář* in *hodinář* "clockmaker").

Ološtiak and Vojteková (2021) delimit three types of compounds according to the type of formants involved. *Proper compounds*[2] are characterized as being composed of two bases, as in (11). *Semi-compounds* are composed of one base and one baseoid (12). Finally, *quasi-compounds* are composed of two baseoids (13).

(11) ***sér|-o-|pozitivní*** ← ***sérum*** ***pozitivní***
seropositive.ADJ serum.NOUN positive.ADJ

(12) ***krypto|politika*** ← ***krypto-*** ***politika***
cryptopolitics.NOUN crypto-.BASEOID politics.NOUN

(13) ***eko|logie*** ← ***eko-*** ***-logie***
ecology.NOUN eco-.BASEOID -logy.BASEOID

[2] The usage of this term by these authors is distinct from Bozděchová's proposal above.

Our conceptualization of neoclassical compounds is mostly congruent with Ološtiak and Vojteková, with a reduction in granularity. Everything the authors consider to be a *baseoid* and some of what the authors consider to be an *affixoid* is considered to be a *neoclassical constituent* (labelled 'neocon' in examples) by us. We also systematically interpret neoclassical constituents as identical whenever their etymology and semantics allow for it, even under circumstances where they undergo formal changes. For instance, the first element of *logografie* 'logography' (*logo-*) and the second element of *sociologie* 'sociology' (*-logie*) are seen to be the same, since they both descend from the same Greek root. In our data, they are represented by the string *-log-*, cf. Section 4.1 for more details.

From the perspective of the parent retrieval task, the simplest case of Czech compounding seems to be compounds formed by simple concatenation of two words, which typically originate in a syntactic phrase and satisfy Bozděchová's definition of *compounding improper*. For instance, the adjective in (14) corresponds directly to the syntactic phrase *vždy zelený* 'always green'. In (15), neither parent word undergoes any morphological change during the compounding procedure, which is characteristic for *compounding improper*, but the resulting noun can be associated with no such phrase, which is typical of *compounding proper*.

(14) ***vždy|zelený*** ← ***vždy*** ***zelený***
evergreen.ADJ always.ADV green.ADJ-NOM.SG

(15) ***garáž|mistr*** ← ***garáž*** ***mistr***
garage supervisor.NOUN garage.NOUN master.NOUN

An interfix is added between the two input words in other compounds, usually *-o-* or *-i-*. This interfix replaces the inflectional ending of any non-final parent; cf. the ending *-a* in the feminine noun *ryba* 'fish' is dropped in (16a). Additionally, stem allomorphy often appears; cf. ∅ ← /e/ in (17).

(16) a. ***ryb|-o-|lov*** ← ***ryba*** ***lov***
fishery.NOUN fish.NOUN hunt.NOUN

b. ***ryb|-o-|lov*** ← ***ryba*** ***lovit***
fishery.NOUN fish.NOUN hunt.VERB

(17) a. ***krv|-o-|tok*** ← ***kre̲v*** ***tok***
bloodflow.NOUN krev.NOUN flow.NOUN

b. ***krv|-o-|tok*** ← ***kre̲v*** ***téci***
bloodflow.NOUN krev.NOUN flow.VERB

Compounding and derivation in one step (18) as well as compounding and conversion in one step (19) are possible, often accompanied by vowel and consonant changes; for instance, in (19) two cases of stem vowel alternation ( ∅ ← /e/ in *ps* ←

*pe̲s* and /o/ ← /e:/ in *vo̲d* ← *vé̲st*), a stem consonant alternation (/d/ ← /s/ in *vod̲* ← *vés̲t*), and an interfix insertion all occur at the same time. Note that in parallel to (19), the compounds in (16a) and (17a) can also be analysed as outputs of compounding and conversion in one step if a noun and a verb are considered as inputs (cf. (16b) and (17b)). In contrast, for *psovod* such an alternative is not available because **vod* is not attestable as a separate noun in Czech. In the data we use in our experiments, both analyses are captured (see Section 4.1).

(18) ***modr|-o-|oký*** ← ***modrý oko***
blue-eyed.ADJ blue.ADJ eye.NOUN

(19) ***ps|-o-|vo̲d*** ← ***pe̲s vé̲st***
dog handler.NOUN dog.NOUN lead.VERB

In (20), the compound is traced back to the noun phrase *chtivý holek* 'wanting of girls', with its original ordering switched. Additionally, there are compounds that cannot be meaningfully split into two parents; cf. the compound in (21) which is composed of a multi-word numeral expression (*dvě a půl* 'two and a half') and the final part which was converted from a noun (*léto* 'year.noun' ← *-letý* '-year.adj').

(20) ***hole̲k|chtivý*** ← ***chtivý holek***
wanting girls.ADJ wanting.ADJ girl.NOUN.GEN.PL

(21) ***dva|a|půl|letý*** ← ***dvě a půl léto***
two-and-a-half-year-old.ADJ two.NUM and.CONJ half.NUM year.ADJ

*Neoclassical compounds*, under our interpretation, constitute what Ološtiak and Vojteková (2021) consider *semi-composition* and *quasi-composition*. The noun *sociologie* 'sociology' in (22) is an example of *quasi-composition* in their framework. In a broader sense, chemical compounds satisfy the definition of semi-composition, as in (23).

Products of reduplication are considered to be compounds for the purposes of this paper, because they formally tend to behave very similarly to compounds (24).[3]

(22) ***soci|-o-|logie*** ← ***-soci- -log-***
sociology.NOUN -soci-.NEOCON -log-.NEOCON

(23) ***tetra|chlor|ethylen*** ← ***-tetra- chlor ethylen***
tetrachlorethylene.NOUN -tetra-.NEOCON chlorine.NOUN ethylene.NOUN

(24) ***čern|-o-|černý*** ← ***černý černý***
pitch black.ADJ black.ADJ black.ADJ

It is worth noting that in spite of all of these formal peculiarities, Czech native speakers tend to find it easy to correctly determine the parents of a given compound. Opportunities for folk etymologies similar to the English *cockroach* (apparently from

[3] Cf. also Hoeksema (2012) who proposes a category of elative compounds.

*cock* + *roach*, actually from the Spanish *cucaracha*) are few and far between. One such example is *medvěd* ('bear') from *med* ('honey') + *jíst* ('eat'), whose etymology is obfuscated by diachronic sound changes. This may lead to a Czech speaker wrongly analyzing the word either as unmotivated or as *med* ('honey') + *vědět* ('know').

## 3. NLP approaches toward word formation

Unlike the long-lasting attention of theoretical linguists, Czech word-formation has come into focus of NLP rather recently. The topic has been addressed primarily by capturing it using static data resources. Additionally, the word formation of other languages has been in the scope of NLP for a much longer time than Czech word formation has.

### 3.1. Derivation trees

Derivancze, which stands for Derivational Analyzer of Czech (Pala and Šmerk, 2015), is a static data resource that can be used to return not only the derivational parents of a given word, but also its derivatives. The tool does not seem to contain compounding relations.

A similar word formation resource for the language, DeriNet, maps derivation by means of linking words to the words they are respectively derived from all the way to their roots, which should canonically be unmotivated. DeriNet has additionally been equipped for handling compounding as well since version 2.0, in that its data format allows for a single lexeme to have multiple parents, and it contains an optional flag for each lexeme signaling whether or not the given lexeme is a compound. Similarly, it is equipped with the possibility of including an *unmotivated* flag (Vidra et al., 2019).

DeriNet version 2.1 (Vidra et al., 2021) contains 33, 938 compounds, of that 2, 691 compounds with linked parents,[4] and a total of words 13, 611 labelled as unmotivated. Furthermore, it contains 664, 430 lexemes which have a single parent, are not roots of a derivation tree, and are lowercase. These items can be assumed to be derivatives or products of conversion.

### 3.2. Compound splitting

Splitting of Czech compounds has been addressed by *Czech Compound Splitter* (Svoboda and Ševčíková, 2021), which is the predecessor of *WFA.ces*. Its primary capability, compound splitting, is parent retrieval limited to confirmed compounds. Analogically, it also performed compound identification, which is word formation classification limited to a binary set of classes – *compounds* and *non-compounds*. The performance and versatility of the tool was what ultimately inspired us to take a new

[4] Manually annotated and added as part of the creation of *Czech Compound Splitter* (Svoboda and Ševčíková, 2021).

direction in word formation analysis and generalize its utility. As there is no other compound splitting tool available for Czech, this task has been demonstrated to be feasible in several other languages.

Henrich and Hinrichs (2011) linked German nominal compounds to their respective parents in GermaNet (Hamp and Feldweg, 1997) using an ensemble of pattern-matching models with an accuracy of 92%. Sugisaki and Tuggener (2018) achieved an F1-score of 92% for finding split-points in German compounds using an unsupervised approach, although they also restricted their efforts to noun-headed compounds only. Ma et al. (2016) achieved an accuracy of 95% using a neural approach trained on the aforementioned GermaNet. Their model performed both splitting and identification of compounds, with the accuracy being an aggregated score of both. Krotova et al. (2020) achieved an accuracy of 96% with a deep-neural model trained on GermaNet data, again restricting themselves to nominal compounds.

A significant amount of research has been dedicated to the study of Sanskrit compounds. This ranges from early, relatively simple rule-and-lexicon based attempts by Huet (2005), who lists no accuracy in his study, to Hellwich and Nehrdich's (2018) deep-learning solution trained on a corpus of $560,000$ Sanskrit sentences with its compound split-points annotated, achieving an accuracy of 96%.

As for other languages, Clouet and Daille (2014) achieved F1-scores of 80% and 63% respectively for finding split-points in English and Russian compounds using a corpus-based statistical approach on manually split compounds.

### 3.3. Stemming

The closest widely used procedural task related to parent retrieval is *stemming*, already mentioned in Section 1. The now classic Porter algorithm was developed in 1979 and published in 1980. There is also a programming language built by Porter, specifically tailored for writing stemmers, called Snowball (Porter, 2001), in which a Czech stemmer called Czech Snowball Stemmer (Chmelař et al., 2011) was implemented.

It has been demonstrated in several languages that NLP tasks such as information retrieval and text classification are significantly improved if the input data is first stemmed. This has been shown for Swedish (Carlberger et al., 2001), Albanian (Biba and Gjati, 2014) and even Czech (Dolamic and Savoy, 2009), which suggests that the task of parent retrieval, addressed in the present paper, might also potentially be of practical interest for the purposes of applications like information retrieval.

Parent retrieval, under our interpretation, differs from stemming in that

- it requires the input to have already been lemmatized;
- it *has to* return a lexical item that appears in the given language's usage as an independent item; and
- it only returns the immediate ancestor of the input word.

For instance, given the English word *unhappiness*, the string **happi* in (25) might be considered to be a correct stemming, despite the fact this string does not occur by itself in written English. When stemming, emphasis is placed on lumping words like *unhappiness*, *happiness* and *happiest* under a single label (**happi* in this case), be it linguistically correct or not. In contrast, (26) or alternatively (27) is what we would expect a parent retriever to do.

(25) ***unhappiness*** ← ****happi***

(26) ***unhappiness*** ← ***unhappy***

(27) ***unhappiness*** ← ***happiness***

Of course, one can use a parent retriever for a purpose similar to that of a stemmer by calling it repeatedly, like in (28) or alternatively (29), which is how a parent retriever can be used for purposes similar to a stemmer. Parent retrieval does *not* handle inflection, so inputting *happiest* into *WFA.ces* may in practice result in unexpected behavior.

(28) ***unhappiness*** ← ***unhappy*** ← *happy*

(29) ***unhappiness*** ← ***unhappy*** ← *happy*

## 4. Data and evaluation methodology

*Word Formation Analyzer for Czech* is a deep-learning based tool, and as such it required data to be trained, tuned, and tested. The following section describes where this data was taken, how it was augmented and preprocessed, and how it was used to fine-tune and test the tool's performance.

### 4.1. Golden data set

The golden data was acquired from DeriNet 2.0 (Vidra et al., 2019). From there, all lexemes that fulfill all of the following requirements at the same time were taken and designated as *derivative*:

- have a single parent,
- are attested in the SYN2015 corpus of Czech (Křen et al., 2016),
- and are not labeled as either *unmotivated* or *compound*,

Then they were paired with their respective DeriNet parent, alongside the class label for *derivative*.

Similarly, all lexemes that fulfilled the following properties were taken and designated as *unmotivated*:

- have no parents,
- are attested in the SYN2015 corpus of Czech,
- and are labeled as *unmotivated*,

The compounds used were compounds from DeriNet with both parents linked. In addition, 285 compounds were hand-annotated specifically as part of creating *WFA.ces*. This data was then compiled into a dataframe of three columns – the first was the lemmas of the lexical items, the second was the parent(s) of these items, and the third contained the respective word class labels.

The data was split into a train set (60%), a test set (20%) and a validation set (20%) according to the *compound* class, as it was the class with the least items. The *unmotivated* and *derivative* classes were split such that there was the same number of items from each of the classes in both the test and validation sets. The rest of the *derivative* items and *unmotivated* items were added into the train set.

Some errors in class labelling were manually found in the test and validation sets, and were appropriately corrected, which resulted in a class imbalance, albeit very slight. The exact composition of the resulting train, test, and validation sets can be viewed in Table 1.

#### 4.1.1. Synthetic data

Because the hand-annotated data set of compounds obtained from DeriNet is too small to reliably train a deep-learning model, we simulated various compound formation procedures that take place in Czech. For example, in (30) we see the process of taking a random adjective stripped of its ending and concatenating it with an *-o-* interfix and with another random adjective. The output is usually nonsensical, but formally correctly formed, like in the example.

| | | | | |
|---|---|---|---|---|
| (30) | ***důležit\|-o-\|neomylný*** | ← | ***důležitý*** | ***neomylný*** |
| | important-infallible.ADJ | | important.ADJ | infallible.ADJ |

For the purposes of training *WFA.ces*, we simulated a number of such compound formation procedures in Python using randomly selected lexemes from DeriNet weighted by their corpus frequency, creating a data set of $280,000$ synthetic compounds. We did not synthesize any derivatives, because the available number of derivative items was deemed sufficient for the purposes of training deep-learning models.

### 4.2. Evaluation methodology

For the purposes of evaluating parent retrieval, we use accuracy, which we define as the proportion of cases wherein *all* parents were correctly predicted by *WFA.ces*.[5]

[5] Parent retrieval accuracy of unmotivated words is equal to the precision of word formation classification, if we consider *unmotivated* to be the positive class.

| Formation class | train | test | validation |
|---|---:|---:|---:|
| Compounds | 1,164 | 284 | 280 |
| Synth. compounds | 280,000 | 0 | 0 |
| Derivatives | 148,921 | 285 | 287 |
| Unmotivated | 4,911 | 284 | 288 |
| Total | 435,280 | 853 | 855 |

*Table 1. The number of lexemes in each formation class, alongside their respective parents, that composed the datasets used to train, develop, and test Word Formation Analyzer for Czech*

In the case of neoclassical compounds, we strictly require the predicted constituents to be correctly hyphenated, as in (31), otherwise the prediction counts as incorrect, cf. (32) and (33).

(31) ***krypt|-o-|fašista*** ← ***-krypt-*** ***fašista*** ✓
cryptofascist.NOUN -crypt-.NEOCON fascist.NOUN

(32) ***krypt|-o-|fašista*** ← ***krypt-*** ***fašista*** ✗
cryptofascist.NOUN crypt-.NEOCON fascist.NOUN

(33) ***krypt|-o-|fašista*** ← ***krypt*** ***fašista*** ✗
cryptofascist.NOUN crypt.NEOCON fascist.NOUN

For the purposes of evaluating *word formation classification*, we rely on convention, using balanced accuracy (balanced so as to compensate for the slightly imbalanced train and validation sets) to assess the model's performance across all three classes; and precision, recall, and F1-score metrics, to evaluate the tool for each word class separately.

For about 38% of the hand-annotated compounds in our dataset, there was ambiguity as to which parents they should be linked to. For instance, *rybolov* 'fishery' may be considered to be either composed of the noun *ryba* 'fish' and the noun *lov* 'hunt', or it alternatively may be analysed as an output of compounding and conversion with the noun *ryba* 'fish' and the verb *lovit* 'to hunt' as inputs (cf. (34a), (34b)). For the purposes of evaluation, both were considered to be correct retrievals. This decision is technical rather than linguistic, and is not supposed to reflect any theoretical preference or view on directionality of conversion and other related issues.

(34) a. ***ryb|-o-|olov*** ← ***ryba*** ***lov*** ✓
fishery.NOUN fish.NOUN hunt.NOUN

(35) b. ***ryb|-o-|lov*** ← ***ryba*** ***lovit*** ✓
fishery.NOUN fish.NOUN hunt.VERB

| Model type | Dropout | Direction | Training iterations |
|---|---:|---|---:|
| default | 0.2 | left to right | 100, 000 |
| transformer | 0.5 | left to right | 900, 000 |
| s2s | 0 | right to left | 30, 000 |

*Table 2. Description of the configurations in the model ensemble used in Word Formation Analyzer for Czech*

## 5. Building and testing the tool

### 5.1. Model ensemble training and tuning

The core of *WFA.ces* was built using the *Marian* framework developed by Junczys-Dowmunt et al. (2018), utilizing an ensemble of three models described in Table 2. All of the models in the ensemble were then trained on the dataset described in Table 1 with layer regularization. The model was trained to take a lexeme from the train set as its input (left-hand side of the arrow in the examples in the previous section) and return its corresponding parent(s) as output (right-hand side of the arrow), separated by spaces if there is more than one parent. The hyperparameters of the model ensemble, such as the dropout rate and number of training iterations, were fine-tuned manually on the test set.

One interesting obstacle that had to be overcome was the fact that, as the FAQ page of the *Marian* project explicitly states:[6] "Convolutional character-level NMT models are not yet supported." Since nothing but isolated lemmas was supported to the model, character-level learning was strictly necessary. We solved this by replacing all spaces (which were only present in the parent sequences of compounds) with an underscore character, and by adding spaces between each character in the string. Thus, *zelenočerný* 'green-black' became *z e l e n o č e r n ý*, and its corresponding parents *zelený černý* became *z e l e n ý _ č e r n ý*. This forced the models to consider each grapheme as a separate word, solving the problem of the models being word-level only.

### 5.2. Tool functioning

*WFA.ces* works by feeding the *Marian* model ensemble an input lexeme $L$ in its lemma form and generating a list of possible parent sequences of size $n$, where $n$ is a natural number chosen by the user. The parent sequences in the list are ordered by their probabilities as predicted by the model ensemble. It then uses simple procedures to find the best candidate in this list to produce the desired outcome for each of the two tasks.

[6] https://marian-nmt.github.io/faq

*Parent retrieval*

*WFA.ces* takes the list of possible parent sequences and uses one of the following reranking procedures, as chosen by the user, to select the best one:

- *First best*: *WFA.ces* simply returns the first parent sequence in the list.
- *Lexicon*: *WFA.ces* uses a provided lexicon to select the first parent sequence in the parent sequence list whose elements are all attestable in that lexicon. If none such sequence can be found in the list, it uses *First best*.
- *Frequency*: *WFA.ces* uses a list of relative corpus frequencies[7] and assigns each element of each sequence in the list of possible parent sequences. It then selects the parent sequence with the smallest sum of squared frequencies.
- *Oracle*: This method is only available if the ground truth is already known, and as such, it is only useful for the purpose of evaluation of the other reranking methods. It returns the correct result, if present in the sequence list.

*Word formation classification*

*WFA.ces* takes the list of possible parent sequences, and:

1. Checks if any of them contains a space character.
2. If yes, it classifies L as a *compound*.
3. If not, it checks whether or not any of the parent sequences are equal to L.
   (a) If yes, it classifies L as an *unmotivated lexeme*.
   (b) If not, it classifies L as a *derivative*.

From this, it follows that when using *WFA.ces* as a *word formation classification* tool, one can consider $n$ to be a user-defined classification threshold: the lower it is, the more *WFA.ces* tends to classify lexemes as *compounds*; the higher it is, the more *WFA.ces* tends to classify words as either *unmotivated* or *derivative*.

### 5.3. Performance evaluation and error analysis

The performance of *WFA.ces* in parent retrieval can be viewed in Table 3. The best reranking method in total is *Lexicon*, though of interest is also *Frequency*, due to its performance in the retrieval of the parents of compounds. This is important, because a user of the tool might decide that the retrieval of compositional parents is more important than the retrieval of derivational parents for the user's purposes, and may select the reranking procedure appropriately. Similarly, a user might decide to use the *First best* method for applications where a reliable lexicon of potential parent words might not be available, such as for the analysis of technical or medical vocabulary, despite the fact that the method exhibits the lowest performance in general performance on our validation set.

In word formation classification, the tool additionally achieved a balanced accuracy of 87% across all three word formation classes. Its performance in this task with

[7] Acquired from DeriNet 2.0 for the purposes of this paper.

| **Lexeme class** | **Reranking method** | | | |
|---|---|---|---|---|
| | Oracle | First best | Lexicon | Frequency |
| Compound | 70% | 56% | 55% | 57% |
| Derivative | 87% | 69% | 75% | 59% |
| Unmotivated | 91% | 71% | 84% | 67% |
| Total | 83% | 65% | 71% | 61% |

*Table 3. The accuracy scores of Word Formation Analyzer for Czech in the task of parent retrieval, broken up for each word formation class, as measured on the validation set for $n = 4$.*

| **Positive lexeme class** | **Classification metric** | | |
|---|---|---|---|
| | Precision | Recall | F1 |
| Compound | 96% | 92% | 94% |
| Derivative | 74% | 97% | 84% |
| Unmotivated | 96% | 70% | 81% |

*Table 4. The Precision, Recall and F1 scores achieved by Word Formation Analyzer for Czech for each word formation class, as measured on the validation set for $n = 4$.*

regards to each class can be viewed in Table 4, wherein each line corresponds to the given class being considered positive and all the others being considered negative for the purposes of the metrics listed in each column. The performance in the classification of compounds is especially promising, suggesting that Czech compounds carry a very distinctive formal fingerprint.

Error analysis confirms that each reranking method presents its own set of strengths and weaknesses. The weakness of the *First best* method is that it often returns strings which are not Czech lemmas (cf. the first line in Table 5). The *Lexicon* method partially solves the problem of nonsensical string outputs, but introduces other problems. For example, it often assumes that neoclassical compounds are unmotivated, because even when a correct splitting comes up in the predicted sequence list, one or more of its constituents might not be present in the lexicon. *WFA.ces* therefore searches for other candidates in the list, wherein the entire neoclassical compound often appears, and is thus returned as the only candidate attestable in the given lexicon (cf. the second line in Table 5). The shortcoming of the *Frequency* reranking, on the other hand, is that it returns highly frequent words even when they are a formally dissimilar candidate from the input (ex. third line in Table 5 – *malý* 'small'). Additionally, the tool has no way of leveraging semantics to its advantage, leading it to analyze *siný* 'light

| Reranking | Input word | Predicted | Correct |
|---|---|---|---|
| *First best* | *plnovous* 'full_beard' | **plnový* | *plný vous* 'full beard' |
| *Lexicon* | *ombrograf* 'ombrograph' | *ombrograf* | *-ombr- -graf-* |
| *Frequency* | *malamut* 'Malamute' | *malý* 'small' | *malamut* 'Malamute' |
| All | *siný* 'light_blue' | *sít* 'sow (verb)' | *siný* 'light_blue' |
| All | *žensky* 'womanly (adv)' | žena 'woman' | ženský 'womanly' |

*Table 5. A sample of the various errors that WFA.ces made in parent retrieval under different reranking methods. Some of the errors were made under all of them.*

blue' as a derivative of *sít* 'to sow' (the penultimate line of Table 5). Some errors were not specific to any particular reranking method. For example, many adverbs in Czech are derived from adjectives. The single most common error in derivational retrieval was in the analysis of such adverbs – instead of retrieving the motivating adjective, *WFA.ces* retrieved the adjective's parent, essentially skipping one derivational step (cf. the last line of Table 5).

## 6. Discussion

In parent retrieval, *WFA.ces* outperforms *Czech Compound Splitter*. Parent retrieval, restricted to compounds, is equivalent to compound splitting; *WFA.ces* exhibits an accuracy of 57% in this task, whereas *Czech Compound Splitter* scores three percentage points less.

The result of *WFA.ces* in word formation classification is somewhat comparable to *Czech Compound Splitters*'s performance of 92% in *compound identification*, but the difference between the two is that the former discriminates between three classes (and thus has a random hit baseline of ca. 33.3%), while the latter discriminates between two classes (having a random hit baseline of 50%). Since the difference between the accuracy scores is five percentage points, but the difference between the baselines is ca. 17 percentage points, we can conclude that *WFA.ces* represents an improvement over *Czech Compound Splitter*. Another feature which sets *WFA.ces* apart in this regard is its classification threshold, which *Czech Compound Splitter* notably lacks, and strongly prefers to identify words as non-compounds.

While *WFA.ces* shows promising results, there is still much to be improved and expanded upon. One of the easiest improvements would be the ability to discriminate between native compounds and neoclassical compounds, since *WFA.ces*'s model ensemble is trained to detect neoclassical constituents by marking them with hyphens. The classification of neoclassical compounds could therefore be implemented without adjusting the deep-learning model ensemble at all. The granularity of this classification could be easily increased even further by discriminating between what

Ološtiak and Vojteková consider to be *semi-compounds* (formed from a neoclassical constituent and a native word) and *quasi-compounds* (formed solely from neoclassical constituents).

Furthermore, products of conversion and derivatives have been grouped into a single class in this study, but it could potentially be valuable to be able to automatically discriminate between the two as well. Since in Czech, conversion is linguistically distinct from derivation by the addition of inflectional affixes as opposed to the addition of derivational affixes, this could hypothetically be achieved by using two lists, one of word formation affixes and another of inflectional affixes. Perhaps the most interesting future development of *WFA.ces* would be its generalization into other languages.

## 7. Conclusions

We presented *Word Formation Analyzer for Czech*, a computational tool for parent retrieval and word formation classification. It is based around an ensemble of deep-learning models built using the *Marian* framework, equipped with output analysis and reranking. It is able to perform word formation classification with 87% balanced accuracy, specifically excelling in discriminating compounds from non-compounds, in which it achieves an F1-score of 94%, and parent retrieval with 71% accuracy, as measured on a separate data set. It outperforms its predecessor, *Czech Compound Splitter*, in every regard. In the future, it would be valuable if *WFA.ces* could be made to distinguish between native and neoclassical compounds, as well as between derivatives and products of conversion. Furthermore, we would like to see the tool generalized into more languages.

## Acknowledgements

This work was supported by the Grant No. GA19-14534S of the Czech Science Foundation, the LINDAT/CLARIAH-CZ project of the Ministry of Education, Youth and Sports of the Czech Republic (project LM2018101), and by 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).

## Bibliography

Biba, Marenglen and Eva Gjati. Boosting text classification through stemming of composite words. In *Recent Advances in Intelligent Informatics*, pages 185–194. Springer, 2014. doi: 10.1007/978-3-319-01778-5_19.

Bozděchová, Ivana. *Tvoření slov skládáním*. Institut sociálních vztahů, Praha, 1997.

Carlberger, Johan, Hercules Dalianis, Martin Duneld, and Ola Knutsson. Improving precision in information retrieval for Swedish using stemming. In *Proceedings of the 13th Nordic Conference of Computational Linguistics (NODALIDA 2001)*, pages 17–22, 2001.

Chmelař, Petr, David Hellebrand, Michal Hrušecký, and Vladimír Bartík. Nalezení slovních kořenů v češtině. In *Znalosti 2011: Sborník příspěvků 10. ročníku konference*, pages 66–77. VŠB-Technical University of Ostrava, 2011. URL `https://www.fit.vut.cz/research/publication/9473`.

Clouet, Elizaveta L. and Béatrice Daille. Splitting of compound terms in non-prototypical compounding languages. In *Workshop on Computational Approaches to Compound Analysis*, pages 11–19, 2014. doi: 10.3115/v1/W14-5702.

Dokulil, Miloš. *Tvoření slov v češtině 1: Teorie odvozování slov*. Academia, Praha, 1962.

Dokulil, Miloš, Karel Horálek, Jiřina Hůrková, Miloslava Knappová, and Jan Petr. *Mluvnice češtiny 1. Fonetika, fonologie, morfonologie a morfematika, tvoření slov*. Academia, Praha, 1986.

Dolamic, Ljiljana and Jacques Savoy. Indexing and stemming approaches for the Czech language. *Information Processing & Management*, 45(6):714–720, 2009. doi: 10.1016/j.ipm.2009.06.001.

Hamp, Birgit and Helmut Feldweg. GermaNet – a lexical-semantic net for German. In *Automatic information extraction and building of lexical semantic resources for NLP applications*, pages 9–15, 1997.

Hellwig, Oliver and Sebastian Nehrdich. Sanskrit word segmentation using character-level recurrent and convolutional neural networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2754–2763, 2018. doi: 10.18653/v1/D18-1295.

Henrich, Verena and Erhard Hinrichs. Determining immediate constituents of compounds in GermaNet. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing 2011*, pages 420–426, 2011.

Hoeksema, Jack. Elative compounds in Dutch: Properties and developments. In *Intensivierungskonzepte bei Adjektiven und Adverben im Sprachenvergleich*, pages 97–142. Kovač Verlag, Hamburg, 2012.

Huet, Gérard. A functional toolkit for morphological and phonological processing, application to a Sanskrit tagger. *Journal of Functional Programming*, 15(4):573–614, 2005. doi: 10.1017/S0956796804005416.

Junczys-Dowmunt, Marcin, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. Marian: Fast Neural Machine Translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, 2018. doi: 10.18653/v1/P18-4020.

Křen, Michal, Václav Cvrček, Tomáš Čapka, Anna Čermáková, Milena Hnátková, Lucie Chlumská, Tomáš Jelínek, Dominika Kovářiková, Vladimír Petkevič, Pavel Procházka, Hana Skoumalová, Michal Škrabal, Petr Truneček, Pavel Vondřička, and Adrian Jan Zasina. SYN2015: Representative Corpus of Contemporary Written Czech. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2522–2528, 2016.

Krotova, Irina, Sergey Aksenov, and Ekaterina Artemova. A Joint Approach to Compound Splitting and Idiomatic Compound Detection. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4410–4417, 2020.

Ma, Jianqiang, Verena Henrich, and Erhard Hinrichs. Letter sequence labeling for compound splitting. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 76–81, 2016. doi: 10.18653/v1/W16-2012.

Ološtiak, Martin and Marta Vojteková. Kompozitnosť a kompozícia: príspevok k charakteristike zloženỳch slov na materiáli západoslovanskỳch jazykov. *Slovo a slovesnost*, 82(2): 95–117, 2021.

Pala, Karel and Pavel Šmerk. Derivancze – derivational analyzer of Czech. In *International Conference on Text, Speech, and Dialogue*, pages 515–523, 2015. doi: 10.1007/978-3-319-24033-6_58.

Porter, Martin F. An algorithm for suffix stripping. *Program: electronic library and information systems*, 14:130–137, 1980. doi: 10.1108/eb046814.

Porter, Martin F. Snowball: A language for stemming algorithms. Published online, October 2001. URL `http://snowball.tartarus.org/texts/introduction.html`. Accessed 21.01.2022, 15.00h.

Sugisaki, Kyoko and Don Tuggener. German compound splitting using the compound productivity of morphemes. In *14th Conference on Natural Language Processing*, pages 141–147, 2018.

Svoboda, Emil and Magda Ševčíková. Splitting and Identifying Czech Compounds: A Pilot Study. In *Proceedings of the Third International Workshop on Resources and Tools for Derivational Morphology* (*DeriMo 2021*), pages 129–138, 2021.

Štícha, František, Miloslav Vondráček, Ivana Kolářová, Jana Bílková, and Ivana Svobodová. *Akademická gramatika spisovné češtiny*. Academia, Praha, 2013.

Štícha, František, Ivana Kolářová, Miloslav Vondráček, Ivana Bozděchová, Jana Bílková, Klára Osolsobě, Pavla Kochová, Zdeňka Opavská, Josef Šimandl, Lucie Kopášková, and Vojtěch Veselý. *Velká akademická gramatika spisovné češtiny 1: Morfologie: Druhy slov / Tvoření slov*. Academia, Praha, 2018.

Valera, Salvador and Alba Ruz. Conversion in English: homonymy, polysemy and paronymy. *English Language and Linguistics*, 25(1):181–204, 2021. doi: 10.1017/S1360674319000546.

Vidra, Jonáš, Zdeněk Žabokrtský, Magda Ševčíková, and Lukáš Kyjánek. DeriNet 2.0: Towards an All-in-One Word-Formation Resource. In *Proceedings of the 2nd Workshop on Resources and Tools for Derivational Morphology*, pages 81–89. Charles University, 2019.

Vidra, Jonáš, Zdeněk Žabokrtský, Lukáš Kyjánek, Magda Ševčíková, Šárka Dohnalová, Emil Svoboda, and Jan Bodnár. DeriNet 2.1, 2021. URL `http://hdl.handle.net/11234/1-3765`. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

**Address for correspondence:**
Emil Svoboda
svoboda@ufal.mff.cuni.cz
Malostranské náměstí 25, 118 01 Praha 1, Czech Republic

The Prague Bulletin of Mathematical Linguistics
NUMBER 118 APRIL 2022

# INSTRUCTIONS FOR AUTHORS

Manuscripts are welcome provided that they have not yet been published elsewhere and that they bring some interesting and new insights contributing to the broad field of computational linguistics in any of its aspects, or of linguistic theory. The submitted articles may be:

- long articles with completed, wide-impact research results both theoretical and practical, and/or new formalisms for linguistic analysis and their implementation and application on linguistic data sets, or
- short or long articles that are abstracts or extracts of Master's and PhD thesis, with the most interesting and/or promising results described. Also
- short or long articles looking forward that base their views on proper and deep analysis of the current situation in various subjects within the field are invited, as well as
- short articles about current advanced research of both theoretical and applied nature, with very specific (and perhaps narrow, but well-defined) target goal in all areas of language and speech processing, to give the opportunity to junior researchers to publish as soon as possible;
- short articles that contain contraversing, polemic or otherwise unusual views, supported by some experimental evidence but not necessarily evaluated in the usual sense are also welcome.

The recommended length of long article is 12–30 pages and of short paper is 6–15 pages.

The copyright of papers accepted for publication remains with the author. The editors reserve the right to make editorial revisions but these revisions and changes have to be approved by the author(s). Book reviews and short book notices are also appreciated.

The manuscripts are reviewed by 2 independent reviewers, at least one of them being a member of the international Editorial Board.

Authors receive a printed copy of the relevant issue of the PBML together with the original pdf files.

The guidelines for the technical shape of the contributions are found on the web site `https://ufal.mff.cuni.cz/pbml`. If there are any technical problems, please contact the editorial staff at `pbml@ufal.mff.cuni.cz`.