Formalisation of the word-formation meaning in language data resources
PhD Thesis Proposal

Lukáš Kyjánek
Charles University
Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics

Abstract

Word-formation processes are used to create new words from the existing ones by changing word-formation meanings. Although there are individual linguistic studies that formalise word-formation meanings, especially in the context of deriving words, these formalisations have not been implemented in the language data resources yet. In order to fill this gap, the proposal brings a review of the current formalisation of the word-formation meanings and deals with labelling and comparing word-formation meanings in the language data resources of several languages.

1 Introduction

One of the ways to name some real-world objects is to create new words by using any of the word-formation processes. An already existing word enters the word-formation process as an input and undergoes changes leading to the creation of the output word, e.g., affixation in *učitel* '*teacher*' → *učitelka* '*female teacher*', conversion in *raněný* '*wounded*' → *raněný* '*wounded man*', compounding in *velký* '*big*' + *město* '*town*' → *velkoměsto* '*city*'. The sum of the input words and the undergone changes is denoted as WORD-FORMATION or STRUCTURAL MEANING (Buzássyová, 1974, p. 31). Note that the changes are not necessarily made in the form of the input word as illustrated in the example of conversion.

Word-formation meanings are more general than LEXICAL MEANINGS, i.e., ways in which speakers use words (Dokulil, 1978, p. 244). The two meanings can match, e.g., the noun *učitelka* '*female teacher*', but their relationship can be also more complicated. In the example of the noun *truhlář* '*cabinetmaker*', the lexical meaning '*a person who makes fine wooden furniture, especially as a job*'[1] is wider that the word-formation meaning '*a person who creates cabinets*' originating from the input noun *truhla* '*cabinet*' and the suffix *-ář* that is shared with other words referring, among other things, to '*a person who does [x]*' in Czech. On other hand, the noun *modřina* '*haematoma*' illustrates the opposite phenomenon – the lexical meaning '*pathol a tumour of clotted or partially clotted blood*' is narrower than the word-formation meaning '*an object which is blue*' originating from the input adjective *modrý* '*blue*' and the suffix *-ina* accompanied by the alternation of *r/ř* and shared with other words referring, among other things, to '*an object which is [x]*' in Czech.

Dokulil (1978) claims the lexical meanings of motivated words are predictable on the basis of the word-formation meanings. However, the crucial assumption of potential research into this predictability is the ability to formalise and predict word-formation meanings first. Štekauer (2005a) exemplifies the formalisation and predictability of word-formation meanings in four experiments, each analysing ten selected English words. Besides his work, the formalisation of word-formation meanings and formal-linguistic features exploitable to predict the meanings are spread to individual linguistic studies focused more on affixation than other word-formation processes.

Although there are dozens of language data resources of word formation (see Kyjánek, 2018), they await the formalisation and prediction of word-formation meanings. The resources have been already harmonised into the same annotation schema in collections Universal Morphology (McCarthy et al., 2020) and Universal Derivations (Kyjánek et al., 2021). They model word-formation processes as relations between lexemes represented by dictionary forms. The resources include and harmonise annotations of lexemes, e.g., part-of-speech categories, but only five resources label a limited set of word-formation meanings. As the current word-formation meanings are formalised and la-

[1] Lexical meanings present in the examples originate from Oxford English Dictionary.

belled inconsistently across the resources, their labels have not been harmonised.

In this proposal and the future dissertation thesis, the task is to find a cross-linguistically applicable approach to the formalisation of word-formation meanings across languages. In addition, the approach should be able to describe word-formation meanings that originate from different word-formation processes. Last but not least, word-formation meanings should be also labelled and compared across languages.

This work focuses on word-formation meanings originating primarily from affixation, conversion, and compounding because of two reasons. First, these three processes are present in most of the fifty-five languages analysed in the typological survey on word formation by Štekauer et al. (2012, p. 309); specifically, compounding is included in 91% of the analysed languages, conversion in 62% languages, and affixation is divided there into suffixation (96%), prefixation (71%), and circumfixation (22%). Second, the three selected processes use different forms to convey word-formation meanings. While affixation and compounding utilise bound and free morphemes respectively, conversion does not use any overt form.

The thesis proposal starts with a review of available approaches to the formalisation of word-formation meanings in affixation, conversion and compounding (Section 2). The language data resources related to word formation, especially those that label word-formation meanings, are described (Section 3). The three tasks of the dissertation topic are discussed separately and illustrated on relevant experiments done in the field. The first task focuses on experiments on the granularity of word-formation meanings and formal-linguistic features utilisable for the prediction of word-formation meanings (Section 4). The second task deals with procedures of labelling word-formation meanings applied to the existing language data resources (Section 5). The third task presents an experiment in which the same word-formation meaning is compared across seven languages (Section 6). In conclusion (Section 7), the near-future perspectives on the topic are provided.

2 Approaches to the formalisation of word-formation meaning

The existing literature primarily deals with the formalisation of the word-formation meanings in affixation as this word-formation process exploits overt affixes to create words. There are also some approaches to the formalisation of word-formation meanings in conversion, which lacks overt affixes, and there is even less research describing the formalisation in compounding or other word-formation processes.

The existing linguistic studies that involve the notion of word-formation meaning are diverse. For instance, the concept of LEXICAL FUNCTIONS (Apresjan et al., 2007, p. 199) from the linguistic framework MEANING–TEXT THEORY by Mel'čuk (1974) embed word-formation meanings, among other phenomena. Lexical functions are defined as triplets of $\{R, X, Y\}$ of (R) certain semantic relation, (X) argument lexeme, and (Y) other lexemes which is the value of R; a set of lexemes is yielded if there are more of such lexemes. Apresjan et al. (2007, p. 209) count around a hundred lexical functions; see the following examples:

S_0(*zkoušet 'to exam')* = *zkouška* '*an exam*',
S_1(*zkoušet 'to exam')* = *zkoušející* '*examiner*',
S_2(*zkoušet 'to exam')* = *zkoušený* '*examinee*'.

The lexical functions can yield more candidates regardless of their word-formation relatedness, e.g., *MAGN(desire) = [strong, keen, intense, fervent, ardent, overwhelming]* as a lexical function for '*a high degree of what is denoted by X*' (Apresjan et al., 2007, pp. 199-200). In addition, lexical functions do not deal exclusively with word formation. Even the three above-mentioned functions are oriented more on the syntactic positions than word formation, so S_1(*učit 'to teach')* = *učitel* '*teacher*' referring to the syntactic position of the agent, whereas S_2(*učit 'to teach')* = *žáci* '*pupils*' referring to the syntactic position of the patient.

The correspondence between the syntactic positions and word-formation meanings conveyed by words assigned to the positions would be an interesting research topic. However, due to their properties, lexical functions are no longer considered to be a good way of formalisation of word-formation meanings in this proposal.

2.1 Word-formation meaning in affixation

The formalisation of word-formation meaning in affixation relies on overt affixes, e.g., the suffix *-ka* in *učitelka* '*female teacher*' motivated by *učitel* '*male teacher*' and referring to female social gender. However, the relation between affixes and

nouns of agents (nomina agentis)
nouns of actors (nomina actoris)
nouns of means (nomina instrumenti)
nouns of results of actions (nomina resultativa)
nouns of bearers of qualities (nomina attributiva)
nouns of beares of a substance relations
nouns of places (nomina loci)
collective and singulative nouns (nomina colectiva et singulativa)
diminutive nouns (diminutiva et meliorativa)
augmentative and pejorative nouns (augmentativa et peiorativa)
nouns of forming sex-opposites (nomina mota)
nouns of young animals
nouns of action (nomina actionis)

Table 1: Word-formation meanings after application of onomasiological theory to Czech nouns by Daneš et al. (1967).

ability	*directional*	*manner*[i]	*relational*
abstraction	*distributive*	*ornative*	*resultative*
action	*durative*	*patient*	*reversative*
agent	*dweller*	*pejorative*	*saturative*[ii]
anticausative	*entity*	*perceptive*	*semelfactive*
augmentative[iii]	*experiencer*	*pluriactionality*	*similative*
causative	*female*	*possessive*	*singulative*
collectivity	*hyperonymy*	*privative*	*state*
comitative	*hyponymy*	*process*	*subitive*
composition	*inceptive*	*purposive*	*terminative*
cumulative	*instrument*	*quality*	*temporal*
desiderative	*iterative*	*reciprocal*	*undergoer*
diminutive[iv]	*location*	*reflexive*	

[i] *viewpoint* [ii] *total* [iii] *ameliorative/intensive* [iv] *attenuative*

Table 2: Comparative semantic concepts (including individual variants in footnotes) for affixation proposed by Bagasheva (2018).

word-formation meanings is many-to-many. The same affix can be used for conveying more word-formation meanings, e.g., *-ka* in *skříňka* '*small cupboard*' ← *skříň* '*cupboard*' for diminution or in *obálka* '*envelope*' ← *obalit* '*to wrap*' for instruments. At the same time, the same word-formation meaning can be conveyed by several formally different affixes, e.g., female social gender by *-ka* in *hráčka* '*female player*' ← *hráč* '*male player*', *-yně* in *ministryně* '*female minister*' ← *ministr* '*male minister*', or *-ová* in *šéfová* '*female boss*' ← *šéf* '*male boss*'.

Dokulil's (1962) ONOMASIOLOGICAL THEORY (later developed on English by Štekauer, 2005b) provides a general theory of word formation. The theory defines NAMING ACTS, i.e., acts of how lexemes are coined in general (Dokulil, 1962, p. 29). The (onomasiological) structure of lexemes consists of a combination of ONOMASIOLOGICAL BASE denoting a class, gender, species, etc., to which the object belongs, and ONOMASIOLOGICAL MARK, i.e., constituents that distinguish the output naming unit from the input naming unit. Their combinations are given by combining ONOMASIOLOGICAL CATEGORIES, namely *substance*, *action*, *quality*, and *circumstance* that correspond to nouns, verbs, adjectives, and adverbs (Štekauer, 2005b, pp. 9–10). For example, the nomina agentis *učitel* '*teacher*' '*a person who teach*' comprises the affix *-tel* (as an onomasiological base specifying the category of substance '*a person*' for which there are also other affixes, such as *-ař/-ář*, *-ce*, *-ec*, *-(n)ík*) and the input word *učit* '*to teach*' (as the onomasiological mark of the category action). In the example of nomina actoris *knihovník* '*librarian*', there is the same onomasiological base but the category of the onomasiological mark changed to the substance as the output word is motivated by the noun *knihovna* '*library*'. Daneš et al. (1967) combined the categories of onomasiological bases and marks and classified Czech nouns created by suffixation to get 13 word-formation meanings, see Table 1. Besides, Dokulil defined three main relations between the onomasiological categories: TRANSPOSITION which changes only part-of-speech category (e.g., *krutost* '*cruelty*' ← *krutý* '*cruel*'), MODIFICATION which slightly changes word-formation meaning by adding new mark (e.g., *klíček* '*small key*' ← *klíč* '*key*'), and MUTATION which is represented by the combination of the base and mark (e.g., *cukřenka* '*sugar bowl*' ← *cukr* '*sugar*').

Concurring with the onomasiological theory and following Haspelmath (2010, p. 663) who supposes comparison of concepts is more adequate for cross-linguistic research than comparison of established language-specific grammatical categories, Bagasheva (2018) elaborates on so-called COMPARATIVE SEMANTIC CONCEPTS that are to be relevant for cross-linguistic research into affixation. Their language independence is grounded in the fundamental concepts of cognition rooted in cognitive linguistics. In comparison to Dokulil's definitions of word-formation meanings, Bagasheva's 51 comparative semantic concepts (see Table 2) are applicable across different types of affixation and across part-of-speech categories. For instance, the noun *psík* '*small dog*' ← *pes* '*dog*' would be labelled *diminutive* as well as the adjective *žlut'oučký* '*yellowish*' ← *žlutý* '*yellow*' and the verb *spinkat* '*to sleep (baby talk)*' ← *spát* '*to sleep*'. The comparative semantic concepts have been already utilised in the cross-linguistic research by Körtvélyessy et al. (2020). They analysed how rich derivational morphology is in 40 European languages on the basis of sets of derivationally-related words for 30 basic words from Swadesh's (1955) core vocabulary.

locatum verbs
location and duration verbs
agent and experiencer verbs
goal and source verbs
instrument verbs
miscellaneous verbs

Table 3: Categories of denominal verbs analysed by Clark and Clark (1979).

2.2 Word-formation meaning in conversion

The formalisation of word-formation meanings in conversion differs from affixation in the fact that there is no overt affix attached to the input lexeme to create a new lexeme. However, conversion is sometimes treated as zero-affixation as it is relatively close to the affixation. The resulting lexeme created by conversion is usable in different contexts and syntactic functions than the respective input lexeme but without changing its form.[2] For example, the masculine verb *bubnovat* '*to drum*' motivated by the noun *buben* '*drum*' in Czech can be easily used as a subject of the sentence (*Buben se protrhl.* '*The drum burst.*') while the verb can be used as a predicate (*Déšť hlasitě bubnoval na střechu.* '*The rain drummed loudly on the roof.*'). To illustrate a difference between the formalisation of word-formation meanings in conversion and affixation, the approach by Clark and Clark (1979) is presented, although it represents only one of many existing studies on conversion.

They focus on what 1,300 English denominal verbs mean in particular contexts. They analysed verbs from different sources, such as newspapers, magazines, novels, etc., and classified them manually. The method of classification is based on paraphrases of verbs, but these paraphrases serve rather as heuristic devices to make groups of verbs with a similar origin (Clark and Clark, 1979, p. 769). As a result, each category shares a general paraphrase to which most of the relevant words fit, e.g., *agent and experiencer verbs* like *to butcher* in '*John butchered the cow.*' can be paraphrased with the agent noun *butcher* as '*John did to the cow the act that one would normally expect [a butcher to do to a cow].*' which works well also for other similar verbs (Clark and Clark, 1979, p. 773). Such an approach led to 6 categories (with potential subdivisions) for denominal verbs, see Table 3. Their labels are taken from *Case grammar* that addresses verbal valency in syntax (Fillmore, 1968, 1971).

[2]In languages with rich inflectional morphology like Czech, formal changes may occur because of removing inflectional endings.

2.3 Word-formation meaning in compounding

The formalisation of word-formation meanings in compounding is a relatively untouched area. This word-formation process connects at least two free morphemes to create a new lexeme, e.g., *velkoměsto* '*city*' ← *velký* '*big*' + *město* '*town*'; and can be accompanied by affixation in the same step, e.g., *dřevorubec* '*lumberjack*' ← *dřebo* '*wood*' + *rubat* '*to mine*' (there is no **rubec* in Czech).

Štekauer (2016) exemplifies the application of the onomasiological theory to compounding. He claims that the onomasiological theory is able to model word-formation meaning but it has a low degree of predictability due to the absence of bound morphemes. However, some of the constituents (free morphemes) may occur frequently in compounds, such as *-man* in *policeman*, *strongman*, *horseman*, *iceman* etc., which might resemble affixation and increase a degree of predictability of such compounds.

As for the traditional semantic classification of compounds, Scalise and Bisetto (2009) defined three relationships between constituents of a compound, namely subordinate, attributive-appositive, and coordinate. Subordinate compounds have a head-complement relation, e.g., *pickpocket* ← *to pick* + *pocket*. Attributive-appositive compounds have a modifier-head relationship, e.g., *bookcase* ← *book* + *case*. Coordinative compounds have a conjunctive relation, e.g., *actor-manager* ← *actor* + *manager*.

3 Language data resources

There are only five resources (two for Czech and one for Croatian, English, and French) with labels of at least some word-formation meanings. They include word-formation meanings labelled for affixation and conversion, but not for compounding.

3.1 CroDeriv

CroDeriv is a manually created lexicon of derivation morphology of Croatian (Filko et al., 2020). It is a database of 14,000 verbs, 1,500 adjectives and 5,500 nouns with an extensive and admirably detailed manual annotation of many phenomena including word-formation meanings. The current inventory includes 21 linguistically-informed labels (see Table 4) concurring with Croatian grammar and reference books but the annotation continues and more labels will be probably introduced (Filko et al., 2020, p. 95). In this language data resource,

action	*literary type*
agent, female	*location*
anatomical part	*number of men involved*
animal, female	*person, both sexes*
deprivation	*plant*
diminutive	*possibility*
disease	*quantity*
drink	*result*
event	*temporal mark*
instrument	*thing*
linguistic term	

Table 4: Labels in CroDeriv.

two approaches to label (a) relations and (b) affixes are used. It allows making visualisations in a form of lexeme-semantic and structure-semantic representations respectively (Filko et al., 2021, p. 121).

3.2 Démonette

Démonette is a language data resource of derivational morphology of French, cf. Hathout and Namer (2014), Hathout and Namer (2016), and Namer and Hathout (2020).[3] During its development, it included several other language data resources of a similar kind, e.g, Glawinette (Hathout and Namer, 2021), making Démonette the largest resource for French. As such, Démonette includes manual or semi-automatically labelled word-formation meaning. The resource models derivationally-related words as derivational paradigms (see Bonami and Paperno, 2018). It labels word-formation meanings on relations in a way inspired by Fillmore (2006)'s Frame Semantics; so far it indicates whether words denote an *action*, an *agent* or a *property* (Sanacore et al., 2019).

3.3 DeriNet

DeriNet is a lexical network of word-formation relations in Czech (Žabokrtský et al., 2016).[4] It is a database of lexemes connected with links corresponding to word-formation relations; each family of derivationally-related words is modelled as an oriented rooted tree (in the Graph Theory terminology) pointing from the base lexeme to the derivative, as understood by (Dokulil, 1962, p. 11). It contains more than one million lexemes extracted from the Czech morphological dictionary MorfFlexCZ (Hajič et al., 2020)[5] and more than 780 thousand derivational relations out of which over 150 thousand relations have been automatically labelled by one of five semantic labels taken from

[3] https://www.demonext.xyz/en/home/

[4] https://ufal.mff.cuni.cz/derinet

[5] MorfFlexCZ contains also lexemes that are not corpus-attested and that are also included in DeriNet.

Label	***Explanation***
k1verb	*process, action or state*
k2pas	*passive participle*
k2rpas	*past passive participle*
k2proc	*active adjectival participles*
k2rakt	*past active adjectival participles*
k2ucel	*objects used for the action*
k1ag	*agent nouns*
k1prop	*property nouns*
k6a	*creation of adverbs from adjectives*
k2pos	*possessive adjectives*
k2rel	*relational adjectives*
k1f	*feminines from general masculines*
k1jmf	*feminines forms of surnames*
k1jmr	*family forms of surnames*
k1obyv	*inhabitant names*
k1dem	*diminutives*
var	*spelling variants (semantic equivalence)*

Table 5: Labels in Derivancze.

Bagasheva's set of comparative semantic concepts, namely *diminutive*, *female*, *possessive*, *iterative*, and *aspect* (Vidra et al., 2019); the label *aspect* is not included in Bagasheva's set.

3.4 Derivancze

Derivacze is a Czech tool that yields a base lexeme and lexemes immediately derived from a lexeme given by the users (Pala and Šmerk, 2015).[6] It contains 255 thousand derivational relations assigned with semantic labels from a set of 17 labels extracted mainly from Czech WordNet (Pala et al., 2011). The labels are represented by technical abbreviations, so Table 5 provides also explanations taken from the documentation of Derivancze. The labels are relatively fine-grained and seem to be limited to a particular part-of-speech category. For example, female counterparts of male professions/types are labelled differently from female counterparts of male surnames, or the label for diminutives is applied only to nouns.

3.5 Morpho-semantic database for English

The Morpho-semantic database is extracted from English WordNet and post-processed by Fellbaum et al. (2007) in such a way that the database includes derivational relations (mostly nominalisations) assigned with original labels from WordNet.[7] Although WordNet focuses on lexical relations between lexemes, such as synonymy, hyperonymy and hyponymy, etc., many language mutations of WordNet include lexemes related through word formation. In the case of this database, there are more than 17 thousand relations with one of 14 labels.

[6] https://nlp.fi.muni.cz/projects/derivancze/

[7] https://wordnet.princeton.edu/download/standoff-files

agent	*material*
body-part	*property*
by-means-of	*result*
destination	*state*
event	*undergoer*
instrument	*uses*
location	*vehicle*

Table 6: Labels in Morpho-semantic database.

4 Experiments on the formalisation of word-formation meaning

This section focuses on data-oriented experiments related to the formalisation of word-formation meanings. In its current state, it deals primarily with word-formation meanings in affixation.

4.1 Granularity of word-formation meanings

Since the thesis presented in this proposal has ambitions to formalise, label and eventually compare word-formation meanings across several languages, Bagasheva's semantic comparative concepts resulting from a typological discussion seem to be a fruitful option. In addition, they have been already partly implemented in DeriNet and the cross-linguistic comparative research into word formation by Körtvélyessy et al. (2020). On the other hand, only word-formation meanings in affixation are captured by the original Bagasheva's work.

The need for a set of labels representing clearly defined word-formation meanings if they are to be compared across languages is also supported by the fact, that even a relatively simple concept like that for female counterparts of male professions or types is defined and labelled differently in the existing studies and data-oriented research. The respective label *female* in the Bagasheva's approach expects only affixation; while the label *social gender* in Bonami and Boyé (2019) includes derivatives and compounds, e.g., *policewoman*; and the label *feminitives* in Nesset et al. (2022) refers to female professionals only and does not include animals, although they are formed by the same means, e.g., *lion* → *lioness* like *actor* → *actress*.

A major complication in setting an adequate level of granularity of word-formation meanings is the fact that the same meaning can be treated as an inflexion in the linguistic tradition of one language while it can be treated as a derivation in another one. Indeed, the borderline between inflexion and derivation seems fuzzy, especially in a cross-linguistic perspective; cf. data-oriented studies by (Bauer, 2004; Štekauer, 2015; Bonami and Paperno, 2018; Bonami and Boyé, 2019). The

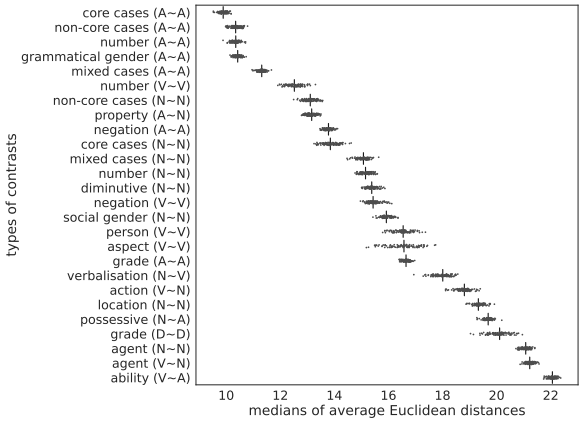

Figure 1: Bootstrap (in 100 iterations) of medians of individual grammatical/word-formation meanings (types of contrasts) on the basis of distributional semantics.

differences in delimiting inflexion and derivation have consequences for the existing language data resources and tools (for example, because of lemmatisation) and for a cross-linguistic comparison of the respective word-formation meaning. For instance, the data-oriented research by Bonami and Boyé (2019) and Mickus et al. (2019) exemplify female counterparts in French as closer to inflectional categories while the experiment on Czech shows they behave similarly to prototypical derivation but are not far from the inflexion (Kyjánek and Bonami, 2022) concurring with the Czech linguistic tradition (Daneš et al., 1967; Štícha et al., 2018). As the language data resources of word formation do not include all word forms but only dictionary forms, it is important to be careful about the linguistic material under analysis.

Fig. 1 shows the results of the mentioned experiment on Czech in which we utilised a vector model of distributional semantics trained on the largest corpus of Czech SYN v9 (Křen et al., 2021). From the Czech language data resources MorfFlexCZ and DeriNet, we extracted word pairs conveying grammatical and word-formation meanings listed on the y-axis. For each of them, we bootstrapped samples of size 200 word pairs with token frequency higher than 50. We calculated average difference vectors for each word pair and measured Euclidean distances between the individual difference vectors and the average difference vector for individual meanings. We then averaged the resulting distances for each sample (points in the graph). If results are plotted, one can see not only the inflexion–derivation scale, where prototypical grammatical meanings like cases are

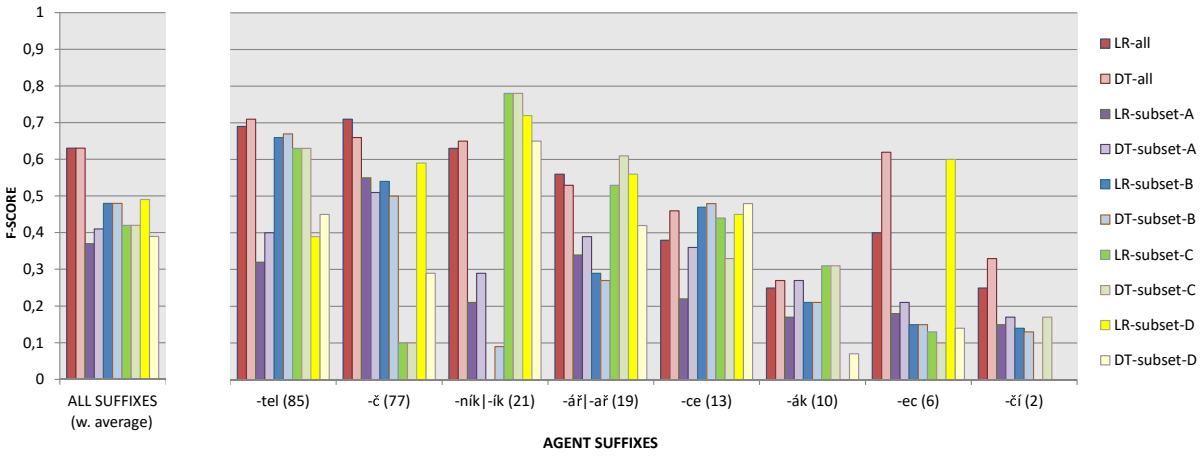

Figure 2: Feature selection to model agent noun formation in Czech. Results for individual suffixes forming agent nouns in Czech (right part of the graph) as well as for all of them aggregated together (left part of the graph). Each sub-set of the analysed features has a different colour; the red one represents a set of all features together. The dark shades of colours represent machine-learning models with the Linear Regression method, while those with light shade represent the Decision Tree method.

on the opposite extreme than prototypical word-formation meanings like agent nouns, but also word-formation meanings standing in the middle of the scale, namely diminution, negation, and female counterparts (labelled as social gender).

Apart from the existing labels of word-formation meanings offered by the reviewed sources, a data-driven approach to exploring word-formation meanings can produce unexpected but interesting results. For instance, Bonami and Naranjo (2023) provide evidence that models of distributional semantics capture word-formation paradigms of a verb and the respective action and agent nouns regardless of the word-formation process used for their creation. This observation opens the possibility of predicting word-formation meanings without the necessity of large training data. And moreover, it can lead to discovering a completely different organisation of word-formation meanings.

4.2 Features of word-formation meanings

The current linguistic research into competition in word formation exploits features that have the potential to a play role in conveying particular word-formation meanings Aronoff (2016). Štekauer (2018, p. 24) claims that each language user is affected by a series of extra-linguistic factors during the process of coining a naming unit. He names the influence of sociolinguistic and psycholinguistic factors, including age, education and profession, bilingual environment etc. However, he also lists prototypical formal-linguistic features affecting how word-formation meanings are conveyed, e.g., phonological factors like the number of syllables of word stem, or morphological factors like the occurrence of a particular affix. The existing studies exploit the mentioned and other formal-linguistic features to research competition in the derivation of verbs from nouns in French (Bonami and Thuilier, 2019), derivation of deverbal nouns in French (Guzmán Naranjo and Bonami, 2022), English suffixes *-ic vs. -ical*, *-ity vs. -ness*, *-ify vs. -ize* (Lindsay, 2012), verb-deriving affixes in English (Plag, 1999), agent nouns in French (Huyghe and Wauquier, 2020, 2021) and English (Lieber and Andreou, 2018), etc.

We also tested a set of formal-linguistic features in our experiment on the formation of agent nouns in Czech, see Ševčíková et al. (2021). The task was to predict an agent affix for a particular verb on the basis of formal-linguistic features. We grouped the features into four different subsets to explore their relevance in the process of formation of agent nouns by affixation:

A. the motivating verb(s): root's final character and theme;
B. the motivating verb(s): number of prefixes, theme, aspect, conjugation class;
C. the derivational paradigm: which motivating items available?, does the verb have a suffixed aspectual counterpart?, does an inanimate homonym exist?;
D. corpus frequency of the motivating items.

We trained two machine-learning models (Decision Tree and Logistic Regression methods) for each of the subsets to observe their relevance for forming agent nouns. Fig. 2 shows that taking all the above-mentioned features into account leads to the best results on average, but there are also interesting results for individual subsets when applied for the prediction of specific affixes. For example, features describing derivational paradigms (subset C) and frequencies (subset D) achieved high results of f-score for predicting agent nouns with the suffixes *-ník*|*-ík*, *-ář*|*-ař*, and *-ec*.

Another option to approach word-formation (and/or lexical) meanings is distributional semantics. The models of distributional semantics define words according to the contexts in which the word occurs (Mikolov et al., 2013). As such, these models may suffer from biases inherited from the training corpora, e.g., Zhou et al. (2019) find out gender bias in distributional models of English and Spanish. However, at the same time, they got an interesting observation that word embeddings of lexemes conveying female social gender in English (e.g., *nurse*) are closer to the female social gender (e.g., *enfemera* '*nurse*') than to the male counterpart (e.g., *enfemero* '*male nurse*') in Spanish.

The key finding is that there are many formal-linguistic features exploitable to formalise individual word-formation meanings. However, the particular word-formation meanings are predictable by a unique set of features.

5 Labelling word-formation meanings in language data resources

The labelling task is complicated because of the homonymy/polyfunctionality of affixes as mentioned above. In the following paragraphs, four labelling procedures applied to the existing language data resources of word formation are presented including their advantages and disadvantages.

One of the methods of labelling is to label word-formation meanings manually. The extraction of words that end with a certain affix (or at least a string similar to the affix or its variants) is relatively easy to do from a given language data resource. However, depending on the size of the data, and the number or diversity of labels, this approach may be inconsistent. Many language data resources are made by this method at the beginning until a substantial amount of data is created to be able to proceed with more advanced methods.

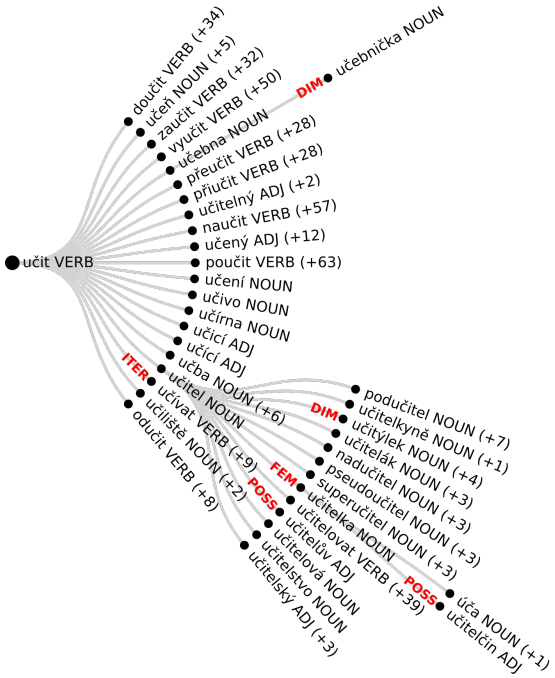

Figure 3: An except of derivationally-related words of the verb *učit* '*to teach*' in DeriNet for Czech. Each relation is labelled by abbreviations of one of the five labels mentioned in the review.

Filko et al. (2021, p. 117) highlight the benefits of a semi-automatic approach to labelling word-formation meanings. They propose to exploit the fact that we can have some expectations regarding the derivation of individual lexemes based on which we can extract and annotate such instances either manually. For example, if female social gender is not labelled yet but we have agent nouns already labelled in Croatian, we can try to find all instances of the subsequent derivation of a verb (*voziti* '*to drive*') → an agent noun (*vozač* '*male driver*') → a noun with feminine grammatical gender and ending with *-ica* or any other affix used for female counterparts (*vozačica* '*female driver*'). This method of labelling has been combined with completely manual annotation in the process of creating CroDeriv. An interesting extension of this method would be to utilise unsupervised machine-learning methods to cluster the analysed instances.

Another method of labelling is to make a precise feature selection and to label word-formation meanings automatically by using machine learning methods. Our initial experiment on labelling word-formation meanings in DeriNet illustrates both the feature selection discussed in Section 4.2 and the labelling procedure applied to DeriNet, see Ševčíková and Kyjánek (2019). We utilised

the existing annotations of almost 18 thousand derivational relations labelled by one of the five word-formation meanings in the digitised language data resource for Czech on which basis we trained and tested the machine-learning model of Multinomial Logistic Regression. The features we used to develop the model were the followings: part-of-speech categories, grammatical genders, grammatical aspects, final character n-grams (bi-, tri-, tetra-, penta-, hexagrams) of both the derivatives and its base word extracted from DeriNet; for adjectival derivatives, we also used a possessivity tag stored in MorfFlexCZ. The classification on the basis of these features achieved outstanding results (more ten 98% of F-score on the testing data set). Fig. 3 shows an excerpt of the labelled data of DeriNet resulting from this experiment.

Once a particular word-formation meaning is labelled at least in one language, it may seem appropriate to utilise the language transfer method to obtain lexemes conveying the same word-formation meaning in another language. We did an experiment heading in that direction by testing ten sources of lexical machine translations to translate female counterparts from Czech DeriNet to six other European languages, namely English, German, Dutch, Russian, French, and Spanish.[8] The preliminary results of machine translation show that there is translation methods suffer from many different aspects. We tested neural network translation by Google Translate but it suffers from hallucination and gender bias. The translations on the basis of the existing multilingual resources, such as Universal WordNet (Melo and Weikum, 2012), PanLex (Kamholz et al., 2014), and CogNet (Batsuren et al., 2019), achieved good precision, but the coverage of vocabulary was unsatisfactory. The translations obtained from custom bilingual dictionaries created on the basis of parallel corpora mitigate the mentioned issues but, on the other hand, they provide too many translation equivalents because of the synonymy of lexemes in those dictionaries. Although providing bad preliminary results, experience from other fields in which language transfer has been used, e.g., development of language data resources of word formation (Vidra and Žabokrtský, 2021) or syntax (Rosa, 2018), indicates that it is worth a try in the labelling word-formation meaning too.

[8]This experiment has not been published yet but it is work-in-progress made in the team cooperation of the START/HUM/010 project.

6 Cross-linguistic comparative research into word-formation meanings

The final thesis should also provide cross-linguistic research into word-formation meanings conveyed across languages either on a micro-level (i.e., affixes or components forming the analysed word-formation meaning) or macro-level (i.e., processes like derivation, compounding etc. forming the analysed word-formation meaning). The future research direction might be also a comparison of formal-linguistic features and their influence on conveying the same word-formation meaning across languages.

This first step towards such cross-linguistic, data-oriented, macro-level morphological research has already been done together with experiments on the labelling method consisting in language transfer presented in the previous section. Our research question – *By what word-formation processes is the same word-formation meaning expressed across several languages?* – have been inspired by Körtvélyessy et al. (2015) who research how new (motivated) lexemes are created in English. We decided to analyse the word-formation meaning of female social gender as we already have this meaning labelled in DeriNet.

We used machine translation to translate 3,746 female counterparts from Czech into other six European languages; we decided to have two Slavic languages (Czech and Russian), two Romance languages (French and Spanish), and three Germanic languages (Dutch, English, and German). Difficulties with the translation methods are described in the previous section. We resolved them by merging and ranking the resulting translations according to a combination of overlaps of translations and the quality of the translation sources. This quality was evaluated manually on a sample of 50 Czech female counterparts translated into other six languages. Means of creation of the translated lexemes, that denote the same entities across languages, were annotated automatically by using Word Formation Analyzer (Svoboda and Ševčíková, 2022),[9] see Table 7. We used several pairwise similarity metrics (Fig. 4) to observe the preliminary results indicating that distributions of the means of creation used for forming female counterparts might correlate with the genetic classification of languages.

Although the pipeline might be fine-tuned, this

[9]The tool has been developed for Czech originally, but we extended it by analysis of the other six languages.

Table 7: Counts of competing means for forming female counterparts in individual languages; their entropy.

Language	non-translated	unmarked	phrase	compound	derivative	unmotivated	Total	Entropy
(CS) Czech	0 (0)	0 (0)	0 (0)	42 (1.1)	3523 (94.0)	181 (4.8)	3746	0.37
(DE) German	476 (12.7)	1119 (29.9)	60 (1.6)	459 (12.2)	1121 (29.9)	511 (13.6)	3746	2.28
(EN) English	116 (3.1)	1952 (52.1)	146 (3.9)	175 (4.7)	912 (24.4)	445 (11.9)	3746	1.90
(ES) Spanish	240 (6.4)	965 (25.8)	77 (2.1)	50 (1.3)	1659 (44.3)	755 (20.1)	3746	1.94
(FR) French	151 (4.0)	1301 (34.7)	111 (3.0)	82 (2.2)	1354 (36.1)	747 (19.9)	3746	1.98
(NL) Dutch	592 (15.8)	919 (24.5)	30 (0.8)	441 (11.8)	835 (22.3)	929 (24.8)	3746	2.32
(RU) Russian	351 (9.4)	579 (15.5)	171 (4.6)	83 (2.2)	2247 (60.0)	315 (8.4)	3746	1.80

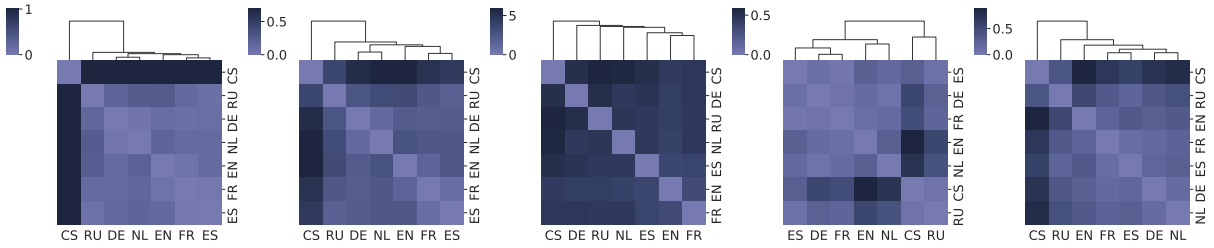

Figure 4: Pairwise similarity metrics used on distributions of the means for forming female counterparts between each pair of languages (from left to right: Kullback-Leibler divergence, Jensen-Shannon divergence, Mutual information, Cosine similarity, Euclidean distance). The order of languages differs across metrics.

so-far unpublished experiment illustrates one of the possible directions of morphological research that would not be possible without the formalisation of word-formation meaning in language data resources. In addition, the presented pipeline might be also utilised for labelling word-formation meanings in under-resourced languages.

7 Conclusion and future perspectives

7.1 Continuation of formalisation

As for the formalisation of word-formation meanings in conversion, the approach by Clark and Clark (1979), which utilises paraphrases to place a lexeme into different contexts, supports the idea of exploiting distributional semantics as its models embed the context of words. Unlike the other formal-linguistic features used in our experiment on agent nouns, a description of derivational paradigms of the lexemes created by conversion (as the conversion is sometimes treated as zero-affixation) might be also useful in the formalisation and further prediction of word-formation meanings.

7.2 Continuation of labelling

Having an idea of the labels and formal-linguistic features applicable in formalisation, we are able to process the labelling. As there are five word-formation meanings already labelled in the Czech DeriNet data resource and there is no data easily extractable from other existing resources to serve as training/testing data in other machine-learning experiments with labelling word-formation meanings in Czech, the possible next step to obtain the data for supervised machine learning experiments might be the semi-automatic procedure proposed by Filko et al. (2021) but extended by the unsupervised learning instead of manual annotations. For instance, the word-formation meaning of *female* is already labelled in DeriNet, so we can extract derivational series comprising a verb or noun → an animate noun with masculine gender → a noun resulting from derivational relation labelled as *female* to obtain *agent nouns*. Since there is a piece of annotation of agent nouns from our previous experiment (Ševčíková et al., 2021), we could enlarge this set. Concurring with Bagasheva (2018) who propose to classify such resulting relations into at least three groups, namely *agent* (i.e., performer of an activity), *dweller* (i.e., an occupant of a specified field), and *patient* (i.e., party to/for whom something is done), we might cluster the extracted data automatically instead of manual annotation. This approach would be challenging because the concepts *dweller* and *patient* are very semantically oriented. The resulting data might serve in the supervised machine-learning experiment on classifying more of the relations conveying any of the three word-formation meanings.

As Bonami and Naranjo (2023) and our experiments above illustrated, the models of distributional semantics may serve not only as features for the formalisation of word-formation meanings

but also as the background for obtaining different organisation of word-formation meanings than expected in the linguistic literature. We could utilise training and testing data from our experiment on the formation of agent nouns (derivational paradigms of agent nouns in Czech) and train a machine-learning model that would predict agent nouns in DeriNet on the basis of distances between lexemes in the respective paradigm.

7.3 Exploration of language transfer

Word-formation meanings labelled in one language might be a base for exploring knowledge transfer into other languages. This knowledge transfer might be useful in both the labelling word-formation meanings and their comparison across languages. For example, having female counterparts labelled in Czech, we might translate them into English to obtain naming units denoting the same entities like we did in the experiment mentioned in Section 6, e.g., *učitelka* '*female teacher*', *herečka* '*actress*', *policistka* '*policewoman*'. As illustrated, such a cross-linguistic approach would have the potential to offer naming units conveying the same word-formation meaning and created by affixation, conversion, compounding or syntactic phrases. However, it is the syntactic phrases that are the technical problem of this idea most of the standard technical solutions would process *female teacher* as two tokens. As a consequence, only the word *female* or *teacher* might be yielded depending on the method used.[10]

The so-far tested methods of lexical machine translation have yielded unsatisfactory results, but there is still possible to try more approaches to the translation. One such way might be to exploit the design of the machine translation systems so sentences containing words conveying the desired word-formation meaning should be translated instead of translating individual words directly. The context in a sentence might help to increase the precision of the resulting translations. Another way might be to test the approach of translating forward and backward as Volker Gast utilised and recommended in his plenary talk in Košice 2022.[11]

[10] In the case of female counterparts, so-called GENERIC MASCULINE WORDS, i.e., words representing both the male and female representatives, may cause troubles and need to be processed separately.

[11] http://kaa.ff.upjs.sk/en/event/43/word-formation-theories-vi-typology-and-universals-in-word-formation-v#toc-plenary-speakers-2

Acknowledgment

Individual parts of this work were supported by Grant No. START/HUM/010 of Grant schemes at Charles University (reg. No. CZ.02.2.69/0.0/0.0/19_073/0016935), Grant No. GA19-14534S of the Czech Science Foundation, and the Charles University Grant Agency (project No. 1176219). It has been using data, tools and services provided by the LINDAT/CLARIAH-CZ Research Infrastructure (https://lindat.cz), supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2018101).

References

Jury D. Apresjan, Igor Boguslavsky, Leonid Iomdin, and Leonid L. Tsinman. 2007. Lexical Functions in Actual NLP-Applications. In Leo Wanner, editor, *Selected Lexical and Grammatical Issues in the Meaning–Text Theory: In honour of Igor Mel'čuk*, page 199–212. John Benjamins Publishing Company.

Mark Aronoff. 2016. Competition and the lexicon. In *Livelli di Analisi e fenomeni di interfaccia. Atti del XLVII congresso internazionale della società di linguistica Italiana*, pages 39–52, Roma. Bulzoni Editore.

Alexandra Bagasheva. 2018. Comparative semantic concepts in affixation. In Juan Santana-Lario and Salvador Valera-Hernández, editors, *Competing Patterns in English Affixation*, pages 33–65. Peter Lang Verlag, Lausanne.

Khuyagbaatar Batsuren, Gabor Bella, and Fausto Giunchiglia. 2019. CogNet: A Large-Scale Cognate Database. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3136–3145, Florence. Association for Computational Linguistics.

Laurie Bauer. 2004. The function of word-formation and the inflection-derivation distinction. In *Words in their Places. A Festschrift for J. Lachlan Mackenzie*, pages 283–292. Vrije Universiteit, Amsterdam.

Olivier Bonami and Gilles Boyé. 2019. Paradigm uniformity and the French gender system. In Matthew Baerman, Oliver Bond, and Andrew Hippisley, editors, *Perspectives on morphology: Papers in honour of Greville G. Corbett*, pages 171–192. Edinburgh University Press.

Olivier Bonami and Matías Guzmán Naranjo. 2023. Distributional evidence for derivational paradigms. In Sven Kotowski and Ingo Plag, editors, *The semantics of derivational morphology: theory, methods, evidence*. De Gruyter, Berlin.

Olivier Bonami and Denis Paperno. 2018. Inflection vs. derivation in a distributional vector space. *Lingue e Linguaggio*, 17:173–195.

Olivier Bonami and Juliette Thuilier. 2019. A statistical approach to rivalry in lexeme formation: French -iser and -ifier. *Word structure*, 12(1):4–41.

Klára Buzássyová. 1974. *Sémantická struktúra slovenských deverbatív*. Veda, Bratislava.

Eve V. Clark and Herbert H. Clark. 1979. When nouns surface as verbs. *Language*, 55(4):767–811.

František Daneš, Miloš Dokulil, Jaroslav Kuchař, et al. 1967. *Tvoření slov v češtině 2: Odvozování podstatných jmen*. Academia, Prague.

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

Miloš Dokulil. 1978. K otázce prediktability lexikálního významu slovotvorně motivovaného slova. *Slovo a slovesnost*, 39(3–4):244–251.

Christiane Fellbaum, Anne Osherson, and Peter E Clark. 2007. Putting Semantics into WordNet's "Morphosemantic" Links. In *Language and Technology Conference*, pages 350–358. Springer.

Matea Filko, Krešimir Šojat, and Vanja Štefanec. 2020. The Design of Croderiv 2.0. *The Prague Bulletin of Mathematical Linguistics*, 115:83–104.

Matea Filko, Krešimir Šojat, and Vanja Štefanec. 2021. Deriving the Graph: Using Affixal Senses for Building Semantic Graphs. In *Proceedings of the Third International Workshop on Resources and Tools for Derivational Morphology (DeriMo 2021)*, pages 120–128. ATILF (CNRS and UNIVERSITÉ DE LORRAINE).

Charles J. Fillmore. 1968. The case for case. In *Universals of linguistic theory*, pages 1–88, New York. Holt, Rinehart and Winston.

Charles J. Fillmore. 1971. Some Problems for Case Grammar. *Working Papers in Linguistics*, 10:245–265.

Charles J. Fillmore. 2006. Frame semantics. In Dirk Geeraerts, editor, *Cognitive linguistics: Basic readings*, pages 373–400. Mouton de Gruyter.

Matías Guzmán Naranjo and Olivier Bonami. 2022. A distributional assessment of rivalry in word formation - supplementary materials. *Zenodo*.

Jan Hajič, Jaroslava Hlaváčová, Marie Mikulová, Milan Straka, and Barbora Štěpánková. 2020. MorfFlex CZ 2.0. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Martin Haspelmath. 2010. Comparative concepts and descriptive categories in crosslinguistic studies. *Language*, 86(3):663–687.

Nabil Hathout and Fiammetta Namer. 2014. Démonette, a French derivational morpho-semantic network. In *Linguistic Issues in Language Technology, Volume 11, 2014 - Theoretical and Computational Morphology: New Trends and Synergies*, volume 11. CSLI Publications.

Nabil Hathout and Fiammetta Namer. 2016. 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'16)*, pages 1084–1091, Portorož. European Language Resources Association (ELRA).

Nabil Hathout and Fiammetta Namer. 2021. Adding Glawinette into Démonette: pratical consequences and theoretical questions. In *Proceedings of the Third International Workshop on Resources and Tools for Derivational Morphology*, pages 70–75, Nancy. ATILF & CLLE, Université de Lorraine.

Richard Huyghe and Marine Wauquier. 2020. What's in an agent? *Morphology*, 30(3):185–218.

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

David Kamholz, Jonathan Pool, and Susan Colowick. 2014. PanLex: Building a Resource for Panlingual Lexical Translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3145–3150, Reykjavik. European Language Resources Association (ELRA).

Michal Křen, Václav Cvrček, Jan Henyš, Milena Hnátková, Tomáš Jelínek, Jan Kocek, Dominika Kováříková, Jan Křivan, Jiří Milička, Vladimír Petkevič, Pavel Procházka, Hana Skoumalová, Jana Šindlerová, and Michal Škrabal. 2021. SYN v9: large corpus of written czech. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Lukáš Kyjánek, Zdeněk Žabokrtský, Jonáš Vidra, and Magda Ševčíková. 2021. Universal Derivations v1.1. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

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

Lukáš Kyjánek and Olivier Bonami. 2022. A Distributional Approach to Inflection vs. Derivation in Czech. In *Word-Formation Theories VI & Typology and Universals in Word-Formation V*, pages 21–22, Košice.

Lívia Körtvélyessy, Alexandra Bagasheva, and Pavol Štekauer, editors. 2020. *Derivational Networks Across Languages*. De Gruyter Mouton, Berlin.

Lívia Körtvélyessy, Pavol Štekauer, and Július Zimmermann. 2015. Word-Formation Strategies: Semantic Transparency vs. Formal Economy. In Laurie Bauer, Lívia Körtvélyessy, and Pavol Štekauer, editors, *Semantics of Complex Words*, pages 85–113. Springer International Publishing, Cham.

Rochelle Lieber and Marios Andreou. 2018. Aspect and modality in the interpretation of deverbal -er nominals in English. *Morphology*, 28(2):187–217.

Mark Lindsay. 2012. Rival suffixes: synonymy, competition, and the emergence of productivity. *Mediterranean Morphology Meetings*, 8:192–203.

Arya D. McCarthy, Christio Kirov, Matteo Grella, Amrit Nidhi, Patrick Xia, Kyle Gorman, Ekaterina Vylomova, Sabrina J. Mielke, Garrett Nicolai, Miika Silfverberg, Tomofey Arkhangelskiy, Nataly Krizhanovsky, Andrew Krizhanovsky, Elena Klyachko, Alexey Sorokin, John Mansfield, Valts Ernštreits, Yuval Pinter, Cassandra L. Jacobs, Ryan Cotterell, Mans Hulden, and David Yarowsky. 2020. UniMorph 3.0: Universal Morphology. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3922–3931, Marseille. European Language Resources Association.

Gerard de Melo and Gerhard Weikum. 2012. Constructing and Utilizing Wordnets using Statistical Methods. *Language Resources and Evaluation*, 46(2):287–311.

Igor A. Mel'čuk. 1974. *Opyt teorii lingvističeskix modelej 'Smysl–Tekst' [An Outline of the Theory of Meaning–Text Type Linguistic Models]*. Nauka, Moscow.

Timothee Mickus, Olivier Bonami, and Denis Paperno. 2019. Distributional Effects of Gender Contrasts Across Categories. In *Proceedings of the Society for Computation in Linguistics (SCiL) 2019*, volume 2, pages 174–184.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Fiammetta Namer and Nabil Hathout. 2020. ParaDis and Démonette – From Theory to Resources for Derivational Paradigms. *The Prague Bulletin of Mathematical Linguistics*, 114:5–34.

Tore Nesset, Alexander Piperski, and Svetlana Sokolova. 2022. Russian feminitives: what can corpus data tell us? *Russian Linguistics*, 46:95–113.

Karel Pala, Tomáš Čapek, Barbora Zajíčková, Dita Bartůšková, Kateřina Kulková, Petra Hoffmannová, Eduard Bejček, Pavel Straňák, and Jan Hajič. 2011. Czech WordNet 1.9 PDT. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Karel Pala and Pavel Šmerk. 2015. Derivancze – Derivational Analyzer of Czech. In *International Conference on Text, Speech, and Dialogue (TSD 2015)*, pages 515–523, Berlin, Heidelberg. Springer Verlag.

Ingo Plag. 1999. On the mechanisms of morphological rivalry: A new look at competing verb-deriving affixes in English. In *Anglistentag 1999 Mainz*, Tübingen. WVT Wissenschaftlicher Verlag Trier.

Rudolf Rosa. 2018. *Discovering the structure of natural language sentences by semi-supervised methods*. PhD dissertation, Charles University, Faculty of Mathematics and Physics.

Daniele Sanacore, Nabil Hathout, and Fiammetta Namer. 2019. Semantic descriptions of French derivational relations in a families-and-paradigms framework. In *Proceedings of the Second International Workshop on Resources and Tools for Derivational Morphology*, pages 15–24, Prague. Charles University.

Sergio Scalise and Antonietta Bisetto. 2009. The classification of compounds. In Rochelle Lieber and Pavol Štekauer, editors, *The Oxford handbook of compounding*, pages 34–53. Oxford University Press, New York.

Emil Svoboda and Magda Ševčíková. 2022. Word Formation Analyzer for Czech: Automatic Parent Retrieval and Classification of Word Formation Processes. *The Prague Bulletin of Mathematical Linguistics*, 118:55–73.

Morris Swadesh. 1955. Chemakum Lexicon Compared with Quileute. *International Journal of American Linguistics*, 21:60–72.

Magda Ševčíková and Lukáš Kyjánek. 2019. Introducing semantic labels into the DeriNet network. *Journal of Linguistics/Jazykovedný časopis*, 70(2):412–423.

Magda Ševčíková, Lukáš Kyjánek, and Barbora Vidová Hladká. 2021. Agent noun formation in Czech: An empirical study on suffix rivalry. In *Second Workshop on Paradigmatic Word Formation Modelling (ParadigMo II)*, page 65, Bordeaux.

Pavol Štekauer. 2005a. *Meaning Predictability in Word Formation: Novel, Context-free Naming Units*. Studies in functional and structural linguistics. John Benjamins Publishing Company.

Pavol Štekauer. 2005b. Onomasiological Approach to Word-Formation. In Pavol Štekauer and Rochelle Lieber, editors, *Handbook of Word-Formation*, pages 207–232. Springer, Dordrecht.

Pavol Štekauer. 2015. The delimitation of derivation and inflection. In Peter O. Müller, Ingeborg Ohnheiser, Susan Olsen, and Franz Rainer, editors, *Word-Formation: An International Handbook of the Languages of Europe*, pages 218–235. De Gruyter Mouton.

Pavol Štekauer. 2016. Compounding from an onomasiological perspective. In Pius ten Hacken, editor, *The Semantics of Compounding*, page 54–68. Cambridge University Press.

Pavol Štekauer. 2018. Competition in Natural Languages. In Juan Santana-Lario and Salvador Valera-Hernández, editors, *Competing Patterns in English Affixation*, pages 15–31. Peter Lang Verlag, Lausanne.

Pavol Štekauer, Salvador Valera, and Lívia Körtvélyessy. 2012. *Word-Formation in the World's Languages: A Typological Survey*. Cambridge University Press, Cambridge.

František Štícha et al. 2018. *Velká akademická gramatika spisovné češtiny 1, Morfologie: Druhy slov / Tvoření slov*. Academia.

Jonáš Vidra and Zdeněk Žabokrtský. 2021. Transferring Word-Formation Networks Between Languages. In *Proceedings of the Third International Workshop on Resources and Tools for Derivational Morphology*, pages 139–148, Nancy. ATILF & CLLE, Université de Lorraine.

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

Pei Zhou, Weijia Shi, Jieyu Zhao, Kuan-Hao Huang, Muhao Chen, Ryan Cotterell, and Kai-Wei Chang. 2019. Examining Gender Bias in Languages with Grammatical Gender. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5276–5284, Hong Kong. Association for Computational Linguistics.

Zdeněk Žabokrtský, Magda Ševčíková, Milan Straka, Jonáš Vidra, and Adéla Limburská. 2016. Merging Data Resources for Inflectional and Derivational Morphology in Czech. In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1307–1314.