

MorfFlex: Handling Rich Morphology

Jaroslava Hlaváčová, Marie Mikulová, Barbora Štěpánková,
Milan Straka, Jan Hajič

Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics
Malostranské náměstí 25, 118 00 Prague 1, Czech Republic
{hlavacova,mikulova,stepankova,straka,hajic}@ufal.mff.cuni.cz

Abstract

We present MorfFlex, a morphological dictionary architecture suitable for languages with extensive regularity in both inflection and derivation. As the primary example of MorfFlex in use we introduce MorfFlex CZ, a morphological dictionary of Czech. It is distributed as a simple, unstructured list of <wordform, lemma, tag> triplets, however, its manually maintained, unpublished source files and conversion scripts encode a sophisticated system of inflectional and derivational patterns. These patterns dramatically reduce the otherwise enormous size of the dictionary, which currently contains over 100 million wordforms and more than 1 million lemmas. The MorfFlex CZ dictionary serves as an essential resource for ensuring the consistency of manual morphological annotation in the Prague Dependency Treebanks and underpins state-of-the-art automatic tools such as MorphoDiTa. In this paper, we focus on: (i) presenting an effective method for managing the rich morphological system within the dictionary, and (ii) demonstrating the utility of such a language resource for maintaining annotation consistency in corpora and supporting the development of advanced NLP applications.

Keywords: morphological dictionary, inflection, derivation, corpus

1. Introduction

Over the course of many years of annotation work, numerous valuable linguistic resources (corpora and dictionaries) have been developed, forming the foundation for a wide range of successful applications and tools in the field of NLP. However, today the future of linguistic annotation lies at the intersection of traditional computational linguistics and modern approaches to natural language processing and artificial intelligence. The role of manual linguistic annotation in application development is steadily diminishing, and annotated resources are required (if at all) only in exceptional cases. This naturally raises the question of whether manual annotation still has a real purpose in the era of large language models. Drawing on several decades of experience with the development of linguistically annotated resources for both theoretical and applied research, we remain firmly convinced that a formalized, reusable source of linguistic knowledge is still a highly desirable goal of computational linguistics.

In this contribution, we present **MorfFlex CZ**,¹ a morphological dictionary of Czech, which is a highly inflectional language. The dictionary was originally developed by Jan Hajič in the late 1980s (Hajič, 2004) as a basis for a spelling checker. Later, it was adapted for commercial applications, particularly lemmatization to support more efficient text search (Hajič and Drozd, 1990). The dictionary then served as a key resource for manual morphological annotation of 1.7 million tokens in the Prague De-

Wordform	Lemma	Tag
<i>létat</i> 'to-fly'	létat	Vf-----A-I--
<i>létám</i> 'I-am-flying'	létat	VB-S---1P-AAI--
<i>nelétám</i> 'I-not-am-flying'	létat	VB-S---1P-NAI--
<i>létáš</i> 'you-are-flying'	létat	VB-S---2P-AAI--
<i>létá</i> 'he-is-flying'	létat	VB-S---3P-AAI--
<i>létáme</i> 'we-are-flying'	létat	VB-P---1P-AAI--
<i>létáte</i> 'you-are-flying'	létat	VB-P---2P-AAI--
<i>nelétají</i> 'they-are-nt-flying'	létat	VB-P---3P-NAI--
<i>nelétají</i> 'they-are-nt-flying'	létat	VB-P---3P-NAI-6
<i>létal</i> 'he-flew'	létat	VpYS----R-AAI--
<i>létali</i> 'they.Masc-flew'	létat	VpMP----R-AAI--
<i>létaly</i> 'they.Fem-flew'	létat	VpTP----R-AAI--
<i>létaly</i> 'they.Neut-flew'	létat	VpNP----R-AAI-6
<i>létala</i> 'she-flew'	létat	VpQW----R-AAI--
<i>létej</i> 'fly!	létat	Vi-S---2--A-I--
<i>létejme</i> 'let's fly!'	létat	Vi-P---1--A-I--
<i>létáno</i> 'flown'	létat	VsNS----X-API--

Table 1: **MorfFlex CZ**: a sample of <wordform, lemma, tag> triplets for the paradigm *létat* 'to-fly'. The complete paradigm contains 106 triplets and 94 unique wordforms (some wordforms are homonymous, e.g. the wordform *létaly* in this sample).

pendency Treebank, a pioneering initiative and one of the very first treebanks and tools ever created. Even before the official release in 2001 (Hajič et al., 2001), an earlier version, annotated at both the morphological and syntactic layers and comprising approximately 400,000 words, was used in 1998 at the Johns Hopkins University workshop. The corpus data provided training material for the first statistical taggers (Hajič et al., 1998) and syntactic parsers (Collins et al., 1999; Charniak, 2000).

¹<https://ufal.mff.cuni.cz/morfflex>

STATISTICS ON MORFFLEX CZ

No. of unique lemmas	1,058,079
No. of <wordform, lemma, tag>	126,906,921

Table 2: Number of unique lemmas and <wordform, lemma, tag> triplets in MorfFlex CZ 2.1.

To date, the dictionary continues to be developed and manually maintained mainly as an essential resource to ensure the consistency of manual morphological annotation in the Prague Dependency Treebanks. The latest version is **MorfFlex CZ 2.1**² (Hajič et al., 2024b), fully compatible with the **Prague Dependency Treebank - Consolidated 2.0** annotation (Hajič et al., 2024a; Mikulová et al., 2026). It also serves as the basis for state-of-the-art automatic tools such as MorphoDiTa (Straková et al., 2014; see Sect. 6).

The architecture of MorfFlex CZ is specially designed for inflectional languages with a large number of endings (suffixes). We refer to this architecture — later applied to other languages with rich inflection (see Sect. 6) — simply as MorfFlex.

In inflective languages, words take endings to mark linguistic cases, grammatical number, gender, tense, etc. Therefore, many wordforms may be related to one lemma. E.g., the Czech word *létat* ('to-fly') can appear as *létám* 'I-am-flying', *nelétám* 'I-am-not-flying',³ *létáš* 'you-are -flying', *letáme* 'we-are-flying', *létejme* 'let's-fly', *létalo* 'it flew', *létali* 'they flew', etc. — there are several tens of wordforms for this type of verb. Corpus-wise (in the Prague Dependency Treebank – Consolidated 2.0 release), there are 230K unique wordforms and 97K lemmas in a corpus of almost 4M words (cf. Table 3). It is therefore crucial to handle endings effectively and to reduce processing costs wherever regularities are found. The dictionary is distributed as a plain unstructured list of the <**wordform, lemma, tag**> triplets (as shown in Table 1). However, the manually maintained, unpublished source files and conversion scripts in fact represent a sophisticated system of inflectional and derivational patterns and rules which reduce the otherwise enormous size of the dictionary. Today, the dictionary contains more than 100 million wordforms and more than 1 million lemmas (cf. Table 2). We release the tool for expanding the source format into the basic <wordform, lemma, tag> triples under an open-source license at <https://github.com/ufal/morfflex-generator>.

In the paper, we describe the Czech morphological dictionary MorfFlex CZ and we focus on presenting:

²<http://hdl.handle.net/11234/1-5833>

³There is also prefix *ne-* 'not/ir/un-', which regularly forms the negative wordforms.

- an effective method for handling rich morphological system of inflectional languages in the dictionary,
- the usefulness of such a language resource for ensuring the consistency of corpus morphological annotation and for developing NLP applications.

The paper is organized as follows. In Sect. 2, the related context of computational morphology is briefly outlined. Sect. 3 provides an introduction to the system of handling rich morphology in MorfFlex. In Sect. 4, the formats of the system is described in more detail: the unpublished source format (Sect. 4.1), the intermediate format (Sect. 4.2), and the basic format (Sect. 4.3) in which the dictionary is distributed. Sect. 5 describes the procedure for converting from the source format to the intermediate format (Sect. 5.1) and then to the basic format (Sect. 5.2). The direct path is described in Sect. 5.3. The use of the dictionary is described in Sect. 6. We conclude in Sect. 7.

CORPUS-WISE STATISTICS

No. of tokens in corpus	3,885,591
No. of unique wordforms	230,448
No. of unique lemmas	97,101

Table 3: Number of unique wordforms and lemmas in the Prague Dependency Treebank - Consolidated 2.0 release.

2. Related Work

In the context of NLP, morphology is primarily employed in the development of tools that recognize relationships between wordforms (morphological analyzers, e.g. *Voikko*,⁴ *Hunmorph* (Trón et al., 2005)) or between related words (derivative tools, e.g. *Sánchez Gutiérrez et al.*, 2017). Many of these tools were originally created for spellchecking purposes (cf. English *SCOWL* (Spell Checker Oriented Word Lists)).⁵ For languages with a rich inflectional structure, for which a comprehensive list would be impractical, a combination of lemmas/roots and inflections is used (cf. *Collatinus*⁶ for Latin, *Ajka*⁷ for Czech, or *Polimorf*⁸ for Polish; for detailed description see e.g. *Crane, 1991, Osolsobě, 1996, Paikens et al.*, 2024). Morphological dictionaries are often created simultaneously with the compilation of a corpus, using manual or semi-manual

⁴<https://voikko.puimula.org/>

⁵<http://wordlist.aspell.net/dicts/>

⁶<https://outils.bibliissima.fr/en/collatinus-web/>

⁷<https://nlp.fi.muni.cz/projects/ajka/>

⁸<https://zil.ipipan.waw.pl/PoliMorf>

annotation, for Slavic languages see e.g. Dobrovoljc et al., 2018 and Ljubešić, 2019. Efforts are also being made to unify morphological features and create a consistent approach to morphology (e.g. *Unimorph*⁹ (Batsuren et al., 2022); *Paralex*¹⁰ (Beniamine et al., 2023)).

3. Morphological Dictionary MorfFlex

The MorfFlex dictionary is a list of <wordform, lemma, tag> triplets. The lemma¹¹ is the basic form of the wordform (usually such a form that is used as an entry in general dictionaries). The tag codes morphological properties of the wordform. The set of all wordforms with the same lemma is called a paradigm. The lemma is usually viewed as a representative of the whole paradigm. For example, the English wordform *fly* belongs to the paradigm represented by the lemma *fly*. The whole paradigm associated with the lemma *fly* is the set {*fly*, *flies*, *flew*, *flying*, *flown*}. Czech paradigms are usually much larger because of the rich inflection of Czech. The paradigm of the Czech equivalent of the English *fly*, namely the verb *létat* has 94 unique wordforms in the MorfFlex CZ dictionary (including 30 rare archaic wordforms; cf. Table 1).

The set of triplets must comply with the so-called *Golden rule of morphology* (Hlaváčová, 2017; Mikulová et al., 2020), which says that any particular pair <lemma, tag> can appear only in one triplet throughout the entire dictionary. In other words, it is not possible that a pair <lemma, tag> can be used for the description of more than one wordform. It guarantees that generating wordforms from <lemma, tag> pairs is unambiguous. We apply tag numbering to distinguish between different types of wordform variants (cf. standard wordform *nelétají* ‘they-are-not-flying’ and non-standard wordform *nelétaj* ‘they-are-not-flying’ in Table 1).

The triplets of the morphological dictionary can be used for generating as well as analyzing wordforms. On the basis of the pair <lemma, tag>, a single wordform is generated, as already explained. The opposite task, analysis, assigns a set of pairs <lemma, tag> to a given wordform. In the latter case, there can be (often are) more such pairs, as the Czech language is massively homonymous. E.g., the morphological analysis assigns two pairs <lemma, tag> to the wordform *létaly* (cf. Table 1).

What we want to present here is that for Czech and other highly inflectional languages with a high degree of regularity (in inflection and/or in derivation), it is possible to describe paradigms using

patterns – sets of rules for the automatic generation of entire paradigms (in the form of triplets). The usage of patterns substantially reduces the size of the dictionary. It is also much more comprehensible for human maintainers.

For various applications, the most convenient format of the morphological dictionary is the format of <wordform, lemma, tag> triplets. We call it the **basic format**. On the other hand, the most convenient format for storing and maintaining the dictionary is the format with patterns, so-called **source format**. Generating the basic format from the source format is not straightforward. In most cases, every record of the source format is first transformed into a set of records of a **virtual intermediate format**, and subsequently into the basic format. This transformation is carried out using rules that are triggered by patterns in both the source format and the intermediate format. The whole procedure is presented as a schema in Fig. 1. The patterns in the source format are **derivational**.¹² They are used for:

1. creating derivations (new words), namely their lemmas
2. assigning inflectional patterns to every derived lemma

The patterns in the intermediate format are **inflectional**. They are used for:

1. generating wordforms associated with lemma
2. assigning the tag describing the morphological properties of every wordform

It is thus not necessary to include some (actually, many) words in the dictionary; they are derived using regular patterns for large sets of words. For instance, there is one record in the source format for the verb *létat* ‘to fly’ (cf. the first row in Table 4). Its derivational pattern (ATN) is “translated” into 20 records of the intermediate format (cf. sample in Table 5). Here, an inflectional pattern is assigned to each of new derived lemmas. This procedure enables not only to create the whole paradigm of the originating verb *létat* ‘to fly’ itself, but also paradigms of the derived words such as *létávat* (‘to use to fly’), *létání* (‘flying’ – noun), *létající* (‘flying’ – adjective), etc. Thus, from the single line in the source format of the verb *létat* ‘to fly’, 3,096 triplets of the basic format are automatically generated (cf. sample in Table 1).

4. MorfFlex Formats

In this section, all three MorfFlex formats are described: source (Sect. 4.1), intermediate (Sect. 4.2) and basic (Sect. 4.3).

⁹<https://unimorph.github.io>

¹⁰<https://www.paralex-standard.org/>

¹¹Though lemma is often considered an abstract object, in MorfFlex it is always expressed as a human readable word.

¹²An exception is formed by the so-called trivial patterns (see Sect. 5.3).

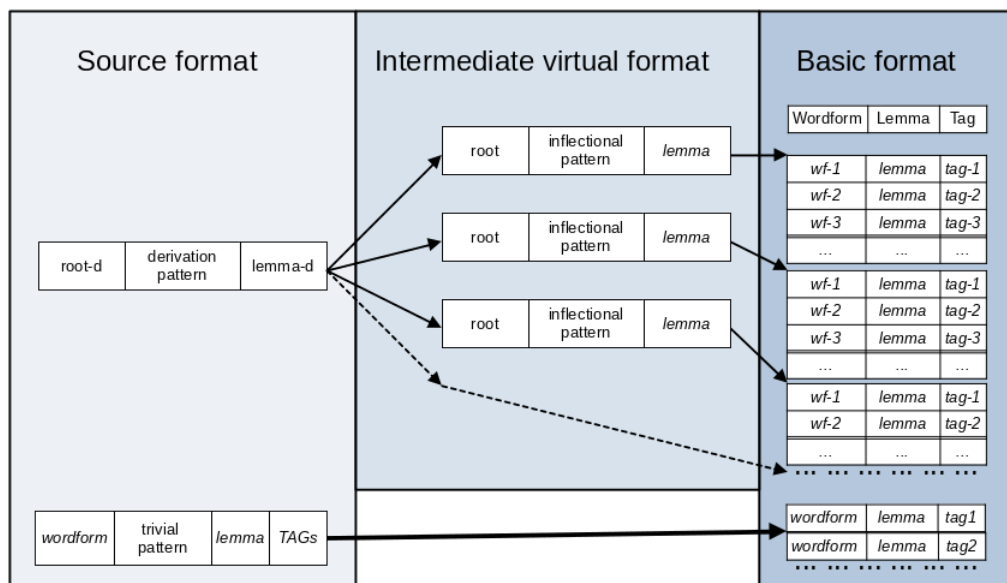


Figure 1: **MorfFlex scheme**. There are three formats: source, intermediate and basic. In the **source format**, two types of records are used. The upper type contains a derivational pattern. According to specific rules for derivational patterns, this record is transformed into one or more records of the **intermediate virtual format**, schematically drawn in the middle column. The derivational pattern is replaced by a set of inflectional patterns, the new roots are created and also the lemma is changed. Each of the records from the intermediate format generates a set of <wordform, lemma, tag> triplets – they are shown in the rightmost part of the scheme, the final **basic format**. The lower part of the scheme shows the simplified procedure for trivial patterns. In that case, the record in the intermediate format would be the same as the source one, which means skipping the intermediate format. The italics font throughout the whole scheme represents the final items (strings), as they are being created while producing the basic format triplets.

4.1. Source format

The source format of MorfFlex is stored in a text format in UTF-8 coding. It enables very easy maintenance in any text editor. Every source format record of the dictionary contains the following pieces of information (prefix d- is for "derivational"):

- **d-root:** beginning of a wordform that does not change within the d-pattern. It does not need to be a grammatical root of a word. The d-root always includes prefixes that appear at the beginning of the wordform, except for the regular negation prefix *ne-* and the superlative prefix *nej-*.
- **d-pattern:** derivational pattern (new lemmas are derived according to the rules stored in this pattern). If it equals to 0 or 0_n, we call it *trivial pattern* and in that case, there must be at least one item TAGs present in the dictionary source line. In this paper, the d-patterns are written in capital letters, contrary to the inflectional patterns.

- **d-lemma:** originating lemma. This item can be enriched with several types of information concerning variants, style, or just an explanation for maintainers. They do not have any influence on the procedure of generating the wordforms and their tags. In case of homonymy, the lemmas are numbered to represent different paradigms, e.g., *jak-1* is for noun ('yak' - animal), *jak-2* is for conjunction ('as'), *jak-3* is for adverb ('how'). See Table 4.
- **TAGs:** morphological tags assigned to the wordform. It is present only when d-pattern is trivial (e.i., equals 0 or 0_n).

Many paradigms are described by more records in the source format. In that case, the d-lemma is identical in all records belonging to that paradigm. Typical examples are verbs with irregular conjugation. Also, irregular or atypical inflection of nouns is described using several records, usually by adding a record with the trivial pattern (cf. record of *ocet* 'vinegar' in Table 4). Examples of the source format records are shown in Table 4.

d-root	d-pattern	d-lemma	TAGs
lét	ATN	létat	
jak	PN18	jak-1	
jak	0	jak-2	J, -----
jak	0	jak-3	Db-----
oc	HD1ET	ocet	
octa	0	ocet	NNIS2-----A---1
octě	0	ocet	NNIS6-----A---1

Table 4: **Record examples in the source format.** The first row is examples of record verb *létat* ‘to-fly’ with derivational pattern. The following 3 rows show records of the homonymous lemma *jak* - one with regular d-pattern; the rest shows the lemma *ocet* ‘vinegar’ with regular d-pattern and two more trivial patterns 0 for the description of two irregular wordforms, not covered by the d-pattern.

i-root	i-pattern	i-lemma	_backlink
lét	atn	létat ‘to fly’	
létajíc	iavg	létající ‘flying - ADJ’	_létat
létání	stn	létání ‘flying - NOUN’	_létat
létáv	atn	létávat ‘to-use-to fly’	_létat
létávání	stn	létávání ‘use-to-fly-NOUN’	_létat

Table 5: Examples of several **records in the intermediate format** derived from the source format record of the d-lemma *létat* ‘to fly’ (cf. the first row in Table 4).

4.2. Intermediate format

The intermediate format is only virtual. It represents a transitional stage in the overall process of generating the final triplets, from the source format to the basic one. The intermediate format records has the same items (except TAGs) as the records in the source format, but their content is different: instead of derivational patterns, there are inflectional patterns. Thus, a virtual intermediate record contains the following pieces of information (prefix i- is for “inflectional”):

- **i-root**: beginning of a wordform that does not change during the inflection within the i-pattern (used for the creation of the whole paradigm for i-lemma)
- **i-pattern**: inflectional pattern.
- **i-lemma**: inflectional lemma – the final lemma that appears in the basic format. The newly derived i-lemmas include a so-called **_backlink** to the originating lemma.

Examples of intermediate format records are shown in Table 5.

4.3. Basic format

The basic format of MorfFlex is a plain unstructured list of <wordform, lemma, tag> triplets:

- **wordform**
- **lemma** - representative wordform. Lemma is unique for the whole paradigm, i.e. a set of wordforms grouped according to inflectional behaviour.
- **tag** - In the positional tag, full inflectional information is coded for the wordform. An overview and explanation of the tag positions are provided in Table 6.

Position	Description
1	Part of speech
2	Detailed part of speech
3	Gender
4	Number
5	Case
6	Possessor’s gender
7	Possessor’s number
8	Person
9	Tense
10	Degree of comparison
11	Negation
12	Voice
13	Verbal aspect
14	Aggregate
15	Variant, style, abbreviation

Table 6: Attributes in positional tags

An example of the final basic format is shown in Table 1. It displays the final triplets generated to the basic format based on the intermediate record shown in the first row of Table 5.

5. MorfFlex Procedure

The procedure for creating a dictionary from the source format is not straightforward; it is carried out in two steps. First, the source format is converted into a virtual intermediate format (Sect. 5.1), from which the basic format is subsequently generated (Sect. 5.2). Those source format records, that contain a trivial pattern (and therefore skip over the intermediate format) are used for generation the triplets of the basic format directly (Sect. 5.3).

5.1. From source to intermediate format

In this section, we describe the procedure of the transformation from the source format (described in Sect. 4.1) into the intermediate format (Sect. 4.2).

Every source format record (except those with trivial patterns) is transformed into one or more

d-root	d-pattern	d-lemma	TAGs
dár	SC1	dárce	

Table 7: Source format record for d-lemma *dárce* ‘donor’.

intermediate records according to the set of special derivational rules. This means that the d-root in the source format can be different from the i-root(s) of the intermediate format, and the d-lemma is changed into a possibly different i-lemma(s). For each new i-root and i-lemma, an inflectional i-pattern is set.

There are one or more derivational rules for each d-pattern. Every derivational rule is represented by a single line that codes the actions to be done with a record of the source format to create the record(s) in the intermediate format. This is a slightly simplified format of a derivational rule:

```
d-pattern r1, r2, r3, r4, r5, r6
```

The explanation is as follows:

- **d-pattern** is a derivational pattern from the source format of the dictionary.
- **r1** is a string that is to be concatenated with the d-root. The result is a new root (i-root). 0 (zero) means an empty string.
- **r2** is an inflectional pattern that is to be used to generate the paradigm of the new lemma.
- **r3** codes the first step of the construction of the new derived lemma (i-lemma).
- **r4** is the ending of the new derived lemma.
- **r5** and **r6** relate to a backlink to the originating d-lemma from which the new i-lemma was created — a derivational link. The derivational link then becomes part of the new lemma, in a form that is rather a note to hint which derivation took place.

The entire procedure of using derivational rules is demonstrated in detail on an example in Sect. 5.1.1.

5.1.1. Example of using a derivational rule

For the demonstration of the derivational rules, we chose the example *dárce* ‘donor’ with the derivational pattern SC1. The source format record is shown in Table 7.

The derivational rules for the (derivational) d-pattern SC1 are in Table 8. Each of the four lines in Table 8 is used to derive one record of the intermediate format as presented in Table 9. It summarizes all the derived lemmas and their inflectional patterns that will generate the paradigm triplets in the basic format. The rows in Table 9 are numbered

	d-pattern	rule
1.	SC1	0,sc1,0,0,0,0
2.	SC1	c,uv,r0,cův,0,0
3.	SC1	ky,ns3n,r0,kyně,0,0
4.	SC1	kynin,in,r0,kynin,r0,kyně

Table 8: Derivational rules for the SC1 d-pattern.

	i-root	i-pattern	i-lemma_back-link
1	dár	sc1	dárce
2	dárc	uv	dárcův_dárce
3	dárky	ns3n	dárkyně_dárce
4	dárkynin	in	dárkynin_dárkyně

Table 9: Records of the intermediate format derived from the source format record presented in Table 7 – according to derivational rules presented in Table 8. The meaning of the i-lemmas are: 1 donor, 2 belonging to donor, 3 donor-woman, 4 belonging to donor-woman.

following the order of the derivational rules in Table 8.

Based on a single line record in the source format (Table 7), four different virtual intermediate records (Table 9) are created according to derivational rules presented in Table 8. Assigned inflectional i-patterns are to be used in the next step (cf. Sect. 5.2.1) for the generation of 250 different triplets in the basic format of the dictionary.

5.2. From intermediate to basic format

The intermediate format of the dictionary (Sect. 4.2) is transformed into the basic format (Sect. 4.3) by means of inflectional i-patterns that appear in records of the intermediate format. The inflectional i-patterns contain a prescription for generating the wordforms and tags of the basic format triplets, while the lemma is simply copied from the intermediate virtual record.

Every i-pattern consists of a set of pairs <ending, tag>, associating the ending (it is simply a string to be concatenated with the i-root) with a particular set of grammatical morphological categories represented by the tag.

The entire procedure of using an inflectional rule is demonstrated on an example in the next Sect. 5.2.1.

5.2.1. Example of using an inflectional rule

To demonstrate the usage of inflectional rules, we chose the intermediate record for the lemma *dárce* ‘donor’ with the i-pattern sc1. The intermediate format record is shown in Table 11 (cf. also the first line of Table 9).

The inflectional rules associated with i-pattern sc1 are demonstrated in Table 10. These rules

sc1			
end	tag	end	tag
-cí	NNMP1-----A----	-cema	NNMP7-----A---6
-cové	NNMP1-----A---1	-ce	NNMS1-----A----
-ců	NNMP2-----A----	-ce	NNMS2-----A----
-cům	NNMP3-----A----	-cí	NNMS3-----A----
-cum	NNMP3-----A---6	-covi	NNMS3-----A---1
-ce	NNMP4-----A----	-ce	NNMS4-----A----
-ci	NNMP5-----A----	-ce	NNMS5-----a----
-cové	NNMP5-----A---1	-ci	NNMS6-----A----
-cích	NNMP6-----A----	-covi	NNMS6-----A---1
-ci	NNMP7-----A----	-cem	NNMS7-----

Table 10: Example of the inflectional *sc1* i-pattern, according to which, for example, the paradigm of the noun *dárce* ‘donor’ is generated.

i-root	i-pattern	i-lemma
dár	sc1	dárce

Table 11: Intermediate record for i-lemma *dárce* ‘donor’.

Wordform	Lemma	Tag
<i>dárci</i> ‘donors-Nominative’	dárce	NNMP1-----A----
<i>dárcové</i> ‘donors-Nomin.’	dárce	NNMP1-----A---1
<i>dárců</i> ‘donors-Genitive’	dárce	NNMP2-----A----
<i>dárcům</i> ‘donors-Dative’	dárce	NNMP3-----A----
<i>dárce</i> ‘donors-Accusative’	dárce	NNMP4-----A----
<i>dárcové</i> ‘donors-Vocative’	dárce	NNMP5-----A----
<i>dárcích</i> ‘donors-Locative’	dárce	NNMP6-----A----
<i>dárci</i> ‘donors-Instrumental’	dárce	NNMP7-----A----

Table 12: A shortened sample of <wordform, lemma, tag> triplets for the paradigm *dárce* ‘donor’.

combine the i-root *dár* with the endings (end) and tags in the Table 10, and generate the resulting basic format triplets (a sample is shown in Table 12).

5.3. Trivial patterns

In the case of trivial patterns, we can interpret the whole procedure as skipping over the intermediate format and generating the basic format triplets directly from the source format (see Fig. 1). There are only two trivial patterns, 0 and 0_n, the latter differs from the former by its ability to generate also a negative wordform, which is formed by placing the prefix *ne-* ‘non/ir/un’ at the beginning of the wordform. No other changes to the root are induced by trivial patterns. Every source record with a trivial pattern must contain the item TAGs, that consists of a set of morphological tags. For each tag, one triplet is generated directly in the basic format. The wordform equals the d-root and the lemma equals the d-lemma.

The trivial patterns are mostly used in case of fully or partially irregular paradigms, but also for uninflected forms. For fully irregular paradigms,

the whole paradigm has to be described as a set of records with the trivial patterns, form by form. Partially irregular paradigms are described by means of non-trivial pattern(s) plus additional record(s) with a trivial pattern (cf. e.g. source format records for the d-lemma *ocet* ‘vinegar’ in Table 4).

The procedure with trivial patterns is demonstrated on an example in the next Sect. 5.3.1.

5.3.1. Examples of using a trivial pattern

To demonstrate the direct procedure with trivial patterns from source format to basic format, we chose the words *dnes* ‘today’ and *zřídka* ‘rarely’. The former one has the trivial pattern 0, the latter one 0_n enabling to create both affirmative and negative wordforms. See source format records in Table 13. The resulting triplets of these records are in Table 14.

d-root	d-pat.	d-lemma	TAGs
dnes	0	dnes	Db-----
zřídka	0 _n	zřídka	Dg-----1@----

Table 13: Source format records of adverbs *dnes* ‘today’ and *zřídka* ‘rarely’ with trivial patterns. The placeholder @ in the second record tag will be replaced with A for the affirmative and N for the negative wordform, see the following Table 14.

Wordform	Lemma	Tag
<i>dnes</i>	dnes	Db-----
<i>zřídka</i>	zřídka	Dg-----1A----
<i>nezřídka</i>	zřídka	Dg-----1N----

Table 14: The resulting <wordform, lemma, tag> basic format triplets for the paradigm *dnes* ‘today’ and *zřídka* ‘rarely’ of the trivial source format records presented in Table 13.

6. Use of MorfFlex

Morphological analysis, part-of-speech tagging are important components of NLP applications. They usually represent initial steps of language processing. Despite recent advances in part-of-speech and morphological tagging, the old truth that more data always gives better results (Banko and Brill, 2001; Church and Mercer, 1993) still holds. At the same time, consistency in data annotation is a very important factor. For morphological annotation, especially for morphologically rich languages with thousands of possible combinations of morphological values, consistency can only be achieved when a dictionary lists all plausible morphological interpretations of all wordforms (cf. Hajič, 2000).

The morphological dictionary MorfFlex CZ has from the very beginning been an essential part of the manual morphological annotation in the Prague Dependency Treebanks (the latest release is **Prague Dependency Treebank – Consolidated 2.0** (Hajič et al., 2024a; Mikulová et al., 2026), containing almost 4M tokens manually annotated for morphology).¹³ As the volume of manually annotated data increases, the dictionary is naturally being expanded and enriched as well. The aim of all annotation projects is to achieve full consistency not only between the data and the dictionary, but also also within both the data and the dictionary themselves (cf. Hajič et al., 2020; Hlaváčová et al., 2019).

The PDT-C corpus and MorfFlex CZ dictionary are used for building models for **MorphoDiTa** (Morphological Dictionary and Tagger),¹⁴ an open-source tool for morphological analysis, generation, tokenization, lemmatization and tagging of texts. It performs morphological analysis and generation using the MorfFlex CZ dictionary. The MorphoDiTa tool achieves state-of-the-art results with a throughput around 10-200K words per second. It has, on average across datasets in PDT-C, F1-scores of 96.27% for tagging and 98.31% for lemmatization (Straka and Straková, 2024). MorphoDiTa is used for automatic POS and morphological annotation of all the Czech corpora available in the Kontext KWIC tool at the LINDAT/CLARIAH-CZ research infrastructure.¹⁵ MorphoDiTa is also one of the two components used for morphological disambiguation of the Czech National Corpus¹⁶ (Petkevič and Jelínek, 2025).

On the manual annotation of PDT-C data, the

¹³For details see a specification of the Czech morphological annotation (Mikulová et al., 2020).

¹⁴<https://lindat.mff.cuni.cz/services/morphodita>

¹⁵<https://lindat.cz>

¹⁶<https://www.korpus.cz>

Czech UDPipe model, is trained. UDPipe¹⁷ is a pipeline for tokenization, tagging, lemmatization and dependency parsing. UDPipe took part in several competitions, reaching excellent results in all of them (Zeman et al., 2018; McCarthy et al., 2019; Sprugnoli et al., 2020). In (Straka and Straková, 2024), the authors show that a model, which combines the deep learning architecture of UDPipe with rescoring by the morphological dictionary MorfFlex (the core of MorphoDiTa), improves over both a deep learning system and a dictionary-based system on their own.

MorfFlex CZ is also aligned with the set of 1,040,126 lexemes contained in **DeriNet** (Vidra et al., 2019), a lexical network which models word-formation relations in the lexicon of Czech.

The MorfFlex framework can also be used for other languages with rich morphology. In addition to Czech, there is also a version for Slovak called **MorfFlex SK** (Hajič and Hric, 2017). This dictionary is also used in MorphoDiTa for the Slovak language.

7. Conclusion

In conclusion, MorfFlex exemplifies how a systematically organized morphological dictionary can effectively manage the complexity of a highly inflective language such as Czech. By combining manually maintained source files with derivational and inflectional patterns, the dictionary achieves a remarkable balance between comprehensiveness and efficiency: from 450K rows in the source format, it produces over 100 million wordforms and more than 1 million lemmas. MorfFlex CZ ensures consistency in manual morphological annotation within the Prague Dependency Treebanks and also provides a robust foundation for state-of-the-art NLP tools like MorphoDiTa.

This contribution highlights the continued relevance of formalized linguistic resources in the era of advanced computational methods. MorfFlex demonstrates that even in highly inflectional languages, careful linguistic modelling can substantially reduce the size of a dictionary while ensuring it remains usable for human annotators and automated systems. Ultimately, this resource underscores the ongoing importance of combining linguistic expertise with computational techniques to advance corpus annotation, NLP development, and the broader study of morphologically rich languages.

¹⁷<https://lindat.mff.cuni.cz/services/udpipe>

8. Acknowledgements

The research and language resource work reported in the paper has been supported by the LINDAT/CLARIN and LINDAT/CLARIAH-CZ projects funded by Ministry of Education, Youth and Sports of the Czech Republic (projects LM2015071, LM2018101, LM2023062). The original annotation has been supported by multiple projects in the past, funded both nationally by the Ministry of Education, Youth and Sports of the Czech Republic and the Czech Science Foundation.

9. Limitations

The presented pattern-based management of rich morphological systems is limited to inflectional languages (primarily Slavic ones). For languages of an agglutinative type, it would need to be adapted. On the one hand, the pattern-based system makes it possible to efficiently generate a huge number of lemmas and wordforms; on the other hand, even a minor change in one pattern can lead to numerous changes in the resulting dictionary, which can be difficult to monitor.

10. Bibliographical References

- Michele Banko and Eric Brill. 2001. [Scaling to Very Very Large Corpora for Natural Language Disambiguation](#). In *Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics*, pages 26–33, Toulouse, France. Association for Computational Linguistics.
- Khuyagbaatar Batsuren, Omer Goldman, Salam Khalifa, Nizar Habash, Witold Kieraś, Gábor Bella, Brian Leonard, Garrett Nicolai, Kyle Gorman, Yustinus Ghanggo Ate, Maria Ryskina, Sabrina Mielke, Elena Budianskaya, Charbel El-Khaissi, Tiago Pimentel, Michael Gasser, William Abbott Lane, Mohit Raj, Matt Coler, Jaime Rafael Montoya Samame, Delio Siconatzi Camaiteri, Esaú Zumaeta Rojas, Didier López Francis, Arturo Oncevay, Juan López Bautista, Gema Celeste Silva Villegas, Lucas Torroba Hennigen, Adam Ek, David Guriel, Peter Dirix, Jean-Philippe Bernardy, Andrey Scherbakov, Aziyana Bayyr-ool, Antonios Anastopoulos, Roberto Zariquiey, Karina Sheifer, Sofya Ganieva, Hilaria Cruz, Ritván Karahóga, Stella Markantonatou, George Pavlidis, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Candy Angulo, Jatayu Baxi, Andrew Krizhanovsky, Natalia Krizhanovskaya, Elizabeth Salesky, Clara Vania, Sardana Ivanova, Jennifer White, Rowan Hall Maudslay, Josef Valvoda, Ran Zmigrod, Paula Czarnowska, Irene Nikkarinen, Aelita Salchak, Brijesh Bhatt, Christopher Straughn, Zoey Liu, Jonathan North Washington, Yuval Pinter, Duygu Ataman, Marcin Wolinski, Totok Suhardijanto, Anna Yablonskaya, Niklas Stoehr, Hossep Dolatian, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Aryaman Arora, Richard J. Hatcher, Ritesh Kumar, Jeremiah Young, Daria Rodionova, Anastasia Yemelina, Taras Andrushko, Igor Marchenko, Polina Mashkovtseva, Alexandra Serova, Emily Prud'hommeaux, Maria Nepomniashchaya, Fausto Giunchiglia, Eleanor Chodroff, Mans Hulden, Miikka Silfverberg, Arya D. McCarthy, David Yarowsky, Ryan Cotterell, Reut Tsarfaty, and Ekaterina Vylomova. 2022. [UniMorph 4.0: Universal Morphology](#). In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 840–855, Marseille, France. European Language Resources Association.
- Sacha Beniamine, Cormac Anderson, Mae Carroll, Matías Guzmán Naranjo, Borja Herce, Matteo Pellegrini, Erich Round, Helen Sims-Williams, and Tiago Tresoldi. 2023. [Paralex: a DeAR standard for rich lexicons of inflected forms](#). Presentation at International Symposium of Morphology.
- Eugene Charniak. 2000. [A Maximum-Entropy-Inspired Parser](#). In *1st Meeting of the North American Chapter of the Association for Computational Linguistics*.
- Kenneth W. Church and Robert L. Mercer. 1993. [Introduction to the Special Issue on Computational Linguistics Using Large Corpora](#). *Computational Linguistics*, 19(1):1–24.
- Michael Collins, Jan Hajič, Lance Ramshaw, and Christoph Tillmann. 1999. [A Statistical Parser for Czech](#). In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, pages 505–512, College Park, Maryland, USA. Association for Computational Linguistics.
- Gregory Crane. 1991. [Generating and Parsing Classical Greek](#). *Literary and Linguistic Computing*, 6(4):243–245.
- Kaja Dobrovoljc, Simon Krek, and Tomaž Erjavec. 2018. [The Sloleks Morphological Lexicon and its Future Development](#). In Vojko Gorjanc, Polona Gantar, Iztok Kosem, and Simon Krek, editors, *Dictionary of Modern Slovene: Problems and Solutions*, pages 42–63. Založba Univerze v Ljubljani, Ljubljana, SI.
- Jan Hajič. 2004. *Disambiguation of Rich Inflection (Computational Morphology of Czech)*. Karolinum, Prague, Czech Republic.

- Jan Hajič, Eduard Bejček, Jaroslava Hlaváčová, Marie Mikulová, Milan Straka, Jan Štěpánek, and Barbora Štěpánková. 2020. [Prague Dependency Treebank - Consolidated 1.0](#). In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5208–5218, Marseille, France. European Language Resources Association.
- Jan Hajič, Barbora Vidová Hladká, Jarmila Panevová, Eva Hajičová, Petr Sgall, and Petr Pajas. 2001. [Prague Dependency Treebank 1.0](#). LDC2001T10, Linguistic Data Consortium, Philadelphia, PA, USA.
- Jan Hajič and Jan Hric. 2017. [MorFlex SK 170914](#). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).
- Jan Hajič. 2000. [Morphological Tagging: Data vs. Dictionaries](#). In *1st Meeting of the North American Chapter of the Association for Computational Linguistics*.
- Jan Hajič, Eric Brill, Michael Collins, Barbora Hladká, Dale Jones, Ching Kuo, Lance Ramshaw, Owen Schwartz, Christoph Tillmann, and Daniel Zeman. 1998. [Core Natural Language Processing Technology Applicable to Multiple Languages](#). Technical report, Center for Language and Speech Processing, Johns Hopkins University, Baltimore, Maryland.
- Jan Hajič and Janus Drozd. 1990. [Spelling-checking for Highly Inflective Languages](#). In *COLING 1990 Volume 3: Papers presented to the 13th International Conference on Computational Linguistics*.
- Jaroslava Hlaváčová. 2017. Golden Rule of Morphology and Variants of Wordforms. *Jazykovedný časopis / Journal of Linguistics*, 68(2):136–144.
- Jaroslava Hlaváčová, Marie Mikulová, Barbora Štěpánková, and Jan Hajič. 2019. Modifications of the Czech morphological dictionary for consistent corpus annotation. *Jazykovedný časopis / Journal of Linguistics*, 70(2):380–389.
- Nikola Ljubešić. 2019. [Inflectional lexicon hrLex 1.3](#). Slovenian language resource repository CLARIN.SI.
- Arya D. McCarthy, Ekaterina Vylomova, Shijie Wu, Chaitanya Malaviya, Lawrence Wolf-Sonkin, Garrett Nicolai, Christo Kirov, Miikka Silfverberg, Sabrina J. Mielke, Jeffrey Heinz, Ryan Cotterell, and Mans Hulden. 2019. [The SIGMORPHON 2019 Shared Task: Morphological Analysis in Context and Cross-Lingual Transfer for Inflection](#). In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–244, Florence, Italy. Association for Computational Linguistics.
- Marie Mikulová, Jiří Mírovský, Milan Straka, Pavlína Synková, Barbora Štěpánková, Jan Štěpánek, and Jan Hajič. 2026. [Prague Dependency Treebank - Consolidated 2.0: Enriching a Complex Annotation Scheme](#). In *Proceedings of the Fifteenth Language Resources and Evaluation Conference*, Palma de Mallorca, Spain. European Language Resources Association.
- Marie Mikulová, Jaroslava Hlaváčová, Jan Hajič, Jiří Hana, Hana Hanová, Barbora Hladká, Barbora Štěpánková, and Daniel Zeman. 2020. [Manual for Morphological Annotation, Revision for the Prague Dependency Treebank - Consolidated 1.0](#). Technical Report TR-2020-64, Institute of Formal and Applied Linguistics, Charles University, Prague, Czech Republic.
- Klára Osolsobě. 1996. *Algoritmický popis české formální morfologie a strojový slovník češtiny*. Disertační práce, Masarykova univerzita, Brno.
- Peteris Paikens, Lauma Pretkalniņa, and Laura Rītuma. 2024. [A Computational Model of Latvian Morphology](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 221–232, Torino, Italia. ELRA and ICCL.
- Vladimír Petkevič and Tomáš Jelínek. 2025. [Automatická morfologická disambiguace korpusů řady syn: spolupráce lingvistické introspekce a strojového učení](#). *Naše řeč*, 108(1):3–40.
- Rachele Sprugnoli, Marco Passarotti, Flavio Masimiliano Cecchini, and Matteo Pellegrini. 2020. [Overview of the EvaLatin 2020 Evaluation Campaign](#). In *Proceedings of 1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 105–110, Marseille, France. European Language Resources Association.
- Milan Straka and Jana Straková. 2024. [Open-Source Web Service with Morphological Dictionary–Supplemented Deep Learning for Morphosyntactic Analysis of Czech](#). In *27th International Conference on Text, Speech and Dialogue*, pages 279–290, Cham, Switzerland. Masaryk University, Springer.
- Jana Straková, Milan Straka, and Jan Hajič. 2014. [Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition](#). In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 13–18, Baltimore, Maryland. Association for Computational Linguistics.

- Claudia Sánchez Gutiérrez, Hugo Mailhot, Hélène Deacon, and Maximiliano Wilson. 2017. [MorphoLex: A derivational morphological database for 70,000 English words](http://link.springer.com/article/10.3758/s13428-017-0981-8). *Behavior Research Methods*, <http://link.springer.com/article/10.3758/s13428-017-0981-8>:1–13.
- Viktor Trón, Gyögy Gyepesi, Péter Halácsky, András Kornai, László Németh, and Dániel Varga. 2005. [Hunmorph: Open Source Word Analysis](#). In *Proceedings of Workshop on Software*, pages 77–85, Ann Arbor, Michigan. Association for Computational Linguistics.
- 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, Czechia. Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics.
- Daniel Zeman, Jan Hajič, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. [CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies](#). In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21, Brussels, Belgium. Association for Computational Linguistics.
- Žabokrtský, Zdeněk. 2024a. [Prague Dependency Treebank - Consolidated 2.0 \(PDT-C 2.0\)](#). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Charles University, Prague, Czech republic. PID <http://hdl.handle.net/11234/1-5813>.
- Hajič, Jan and Hlaváčová, Jaroslava and Marie Mikulová and Milan Straka and Barbora Štěpánková. 2024b. [MorfFlex CZ 2.1](#). LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Charles University, Prague, Czech republic. PID <http://hdl.handle.net/11234/1-5833>.

11. Language Resource References

- Hajič, Jan and Bejček, Eduard and Bémová, Alevtina and Buráňová, Eva and Fučíková, Eva and Hajičová, Eva and Havelka, Jiří and Hlaváčová, Jaroslava and Homola, Petr and Ircing, Pavel and Kárník, Jiří and Kettnerová, Václava and Klyueva, Natalia and Kolářová, Veronika and Kučová, Lucie and Lopatková, Markéta and Mareček, David and Mikulová, Marie and Mírovský, Jiří and Nedoluzhko, Anna and Novák, Michal and Pajas, Petr and Panevová, Jarmila and Peterek, Nino and Poláková, Lucie and Popel, Martin and Popelka, Jan and Romportl, Jan and Rysová, Magdaléna and Semecký, Jiří and Sgall, Petr and Spoustová, Johanka and Straka, Milan and Straňák, Pavel and Synková, Pavlína and Ševčíková, Magda and Šindlerová, Jana and Štěpánek, Jan and Štěpánková, Barbora and Toman, Josef and Uřešová, Zdeňka and Vidová Hladká, Barbora and Zeman, Daniel and Zikánová, Šárka and