EURALEX XIX

Congress of the European Association for Lexicography

Lexicography for inclusion

7-11 September 2021

Ramada Plaza Thraki
Alexandroupolis, Greece

www.euralex2020.gr

Proceedings Book
Volume 1

Edited by Zoe Gavriilidou, Maria Mitsiaki, Asimakis Fliatouras

EURALEX Proceedings

ISSN 2521-7100

ISBN 978-618-85138-1-5

Edited by: Zoe Gavriilidou, Maria Mitsiaki, Asimakis Fliatouras

English Language Proofreading: Lydia Mitits and Spyridon Kiosses

Technical Editor: Kyriakos Zagliveris

2020 Edition

The MorfFlex Dictionary of Czech as a Source of Linguistic Data

Štěpánková B., Mikulová M. , Hajič J.

Charles University, Prague, Czech Republic

Abstract

In this paper we describe MorfFlex, the Morphological Dictionary of Czech, as an invaluable resource for exploring the formal behavior of words. We demonstrate that MorfFlex provides valuable and rich data allowing to elaborate on various morphological issues in depth, which is also connected with the fact that the MorfFlex dictionary includes words throughout the whole vocabulary range, including non-standard units, proper nouns, abbreviations, etc. Moreover, in comparison with typical monolingual dictionaries of Czech, MorfFlex also captures non-standard wordforms, which is very important for Czech as a language with a rich inflection. In the paper we also demonstrate how particular information on lemmas and wordforms (e.g. variants, homonymy, style information) is marked and structured. The dictionary is provided as a digital open access source available to all scholars via the LINDAT/CLARIAH-CZ language resource repository. It is available in an electronic format, and also in a more human-readable, browsable and partly searchable form.

Keywords: Morphology; Dictionary; Czech; Lemma; Wordform; Tag

1 The Morphological Dictionary of Czech: MorfFlex

MorfFlex (Hajič et al. 2020a) is a dictionary of Czech wordforms with detailed morphological information (see a more detailed description in Hajič 2004, Hlaváčová et al. 2019, Mikulová et al. 2020). The MorfFlex dictionary represents more than 100 million wordforms and more than 1 million lemmas. It has been developed gradually since 1988 and the latest electronic version will be published in 2020 (Hajič et al. 2020a). It is also available in a more human-readable, browsable and partly searchable form via online MorphoDita tool.[1]
MorfFlex has the following goals, namely providing:

- a basis for consistent morphological annotation of the Prague Dependency Treebanks (see more Hajič et al. 2017, Hajič et al. 2020b) which serve as a training data for various NLP tasks (tagging and lemmatization, cf. the tool Morphodita (Straková et al. 2014));
- a basis for tagging and lemmatization of other synchronic corpora of Czech, e.g. the Czech National Corpus (Hnátková et al. 2011) and the web corpora Araneum (Benko 2014);
- a resource for linguistically-oriented research, particularly for describing morphological characteristics of Czech.

In this paper, we concentrate on the last aspect.

2 MorfFlex as a Resource for Linguistic Research

In typical monolingual dictionaries, which focus on the meaning of lexical units, there is usually a brief morphological description followed by the definition of the meaning and examples for each word. In current Czech dictionaries (e.g. *Slovník spisovného jazyka českého*, *Slovník spisovné češtiny*) the morphological description contains the part of speech, gender (for nouns), grammatical aspect (for verbs), and typically one or more supporting inflectional suffixes (endings), mostly genitive singular ending for nouns and 1st singular present and 2nd singular imperative endings for verbs, which help to identify, for a speaker of Czech, the complete inflectional paradigm. The macrostructure as well as the microstructure of MorfFlex is completely different, since it specializes in morphology. MorfFlex is not a dictionary of words, but of wordforms. Each entry is represented by a triple composed of a wordform, lemma, and tag. Wordforms are organized into paradigms according to their formal morphological behavior. The paradigm is identified by a unique lemma. For each wordform, full inflectional information is encoded in a tag. We can search the dictionary by lemma or by wordform.
The morphological system plays an important role in Czech, which is, like other Slavic languages, highly inflected. As the results from the European Survey of Dictionary Use and Culture (cf. Kosem et al. 2019, the Czech Republic local dataset) confirm, grammatical information is one of the pieces of information most searched for by the users of Czech monolingual dictionaries.
This fact probably reflects the Czech language situation. Bermel (2000) describes it as quasi-diglossic, i.e. a situation which is characterized by the existence of two varieties used by a single language community.[2] Besides the main variety representing Standard Czech, the other variety is also significant for Czech – it is a non-standard variety, which covers

[1] http://lindat.mff.cuni.cz/services/morphodita/run.php
[2] Bermel follows Ferguson's description of diglossia (1959), specifying the term quasi-diglossia: "In contrast to classic diglossic system, the high and low codes are mutually comprehensible." (Bermel 2007: 51).

most of the Czech language area, mainly of Bohemia.[3] This variant is used mainly in spoken informal communication, the so called *obecná čeština* (*Common Czech*, sometimes also *Colloquial Czech*; cf. Hoffmannová 2013). This variety is present in both the lexicon (lemmas) and the morphology (wordforms), e.g. most adjectives have a complete paradigm of Common Czech endings (cf. also examples 1-4 below). While Czech monolingual dictionaries are traditionally focused on description of standard Czech, MorfFlex captures the morphology (and to a lesser extent the lexicon) of Standard Czech, Common Czech and to some extent some dialects.

(1) lemma: *kývat* (standard) vs. *kejvat* (Common Czech) [infinitive: *'to sway'*]
(2) lemma: *okno* (standard) vs. *vokno* (Common Czech) [nom. sg. neuter: *'window'*]
(3) wordform: *kývají* (standard) vs. *kejvaj* (Common Czech) [3rd pl. present: *'sway'*][4]
(4) wordform: *mladých* (standard) vs. *mladejch* (Common Czech) [loc. pl. adj.: *'young'*]

3 Generation: From Lemma to its Wordforms

MorfFlex covers all possible types of words (tokens) that occur in real Czech texts, i.e. Czech words, loan words, foreign words, proper nouns, abbreviations, parts of words, and numerals. For each lemma, the following information is captured:

- paradigm
- stylistic characteristics
- homonymy
- semantic labels
- derivative relation

Unlike paradigms which are represented by a set of wordforms and tags, the labels and indexes are not a part of the tag but refer to the whole lemma. Note that in MorfFlex, neither the semantic label, nor the stylistic characteristic, not the homonymy indexing are transposed from any Czech monolingual dictionary, but they are provided by manual annotation (cf. Hajič et al. 2020b).

3.1 Paradigms Comprising all Wordforms, including Non-standard Ones

MorfFlex captures both the singular and the plural set of wordforms of all inflected words, even of proper nouns. As mentioned above, MorfFlex is not only focused on Standard Czech, therefore the paradigms also provide non-standard variants and capture the stylistic characteristics of wordforms. The distinction between standard and non-standard wordforms is captured by a number on the last, 15th position in the tag (see Sect. 4.1.3 for another use of the 15th position in the tag). No value (in the tag) indicates the primary wordform, numbers 1-5 mark standard variants, and numbers 6-9 non-standard variants. See Table 1, where variant wordforms of the instrumental plural of the name *Thales* are shown; the first two variants belong to the standard variety, the last wordform represents the non-standard one.

Thalesi	Thales_;Y	NNMP7-----A----
Thalety	Thales_;Y	NNMP7-----A---1
Thalesema	Thales_;Y	NNMP7-----A---6

Table 1: Standard and non-standard variants.

3.2 Homonymy of Lemmas

Unlike typical monolingual dictionaries, MorfFlex does not capture any differences in meanings of homonymous words;[5] it however distinguishes lemmas with the same spelling but different formal morphological behavior. Each homonymous lemma is marked by an index, e.g. *drát-1* (the noun *'wire'*), *drát-2* (the verb *'to pluck'*). In some cases, however, essential syntactic characteristics are taken into account: for example, homonymous forms of uninflected words are considered to be different, and therefore represented by two lemmas, e.g. *přece* which is in accordance to its behavior/function in a sentence interpreted as a conjunction *'despite'* (lemma *přece-1*) or as a particle *'after all'* (lemma: *přece-2*), and it has two lemmas with different indexes and different tags in the dictionary.

3.3 Stylistic Characteristics

Although MorfFlex is primarily focused on morphology, in certain cases the stylistic characteristization of a word (lemma) is also provided. This information is consistently marked for variants which have the same declension or conjugation but different stylistic characteristics. In such cases, one lemma is selected as the basic one and the others are

[3] Sometimes this variety is classified as an interdialect (Šipková 2017).
[4] Often various combinations of standard and Common Czech are possible, e.g. *kývají, kývaj, kejvají, kejvaj*.
[5] Thus, in MorfFlex we do not distinguish e.g. the feminine noun *matka* (*'mother'*) from *matka* (*'nut'*), even though the former is animate and specific possessive forms can be derived.

marked and linked to it. Still, the rule applies that meanings of words are not taken into account. We use a set of labels for distinguishing standard and non-standard variants. (See Table 2.) In examples 5-7 several variant lemmas are mentioned and the way they are linked is shown: two standard variants of the name *Thalés* and *Thales,* the noun *býk 'bull'* and its non-standard, colloquial variant *bejk*, and the noun *večer* '*evening*' and its dialect variant *večír*.

(5) *Thalés_;Y_,s_^(^DD**Thales) → Thales_;Y*
(6) *bejk_,h_^(^GC**býk) → býk*
(7) *večír-1_,n_^(^GC**večer-1) → večer-1*

Stylistic characteristics		Label
standard	literary	s
	archaic	a
non-standard	dialectal	n
	non-standard, Common Czech	h
	expressive	e
	slang, argot, cant	l
	offensive, vulgar	v

Table 2: List of stylistic labels.

3.4 Semantic Labels

Some nouns are also marked by the so-called *semantic label* (see Table 3), i.e. a label (or more labels) classifying them into a particular semantic group. (See Table 3 for the list of the semantic labels.) It is primarily used for nouns starting with a capital letter, both Czech ones and those integrated into the Czech morphological system. For example, in Table 1, the label Y following the lemma *Thales* indicates a person name. Semantic labels help to tell homonymous words apart – e.g., they distinguish the animateness of nouns, as e.g. in *McIntosh-1_;Y* the Y refers to the animate behavior of the word, and *G* and *m* in *McIntosh-2_;G_;m* to its inanimate behavior. Semantic labels also serve to verify that the first capital letter is used properly, i.e. all noun lemmas starting with a capital letter have to be marked by a semantic label.

Code	Definition
Y	person names (given, family, etc.)
E	nationalities, citizen, ethnic, and other named groups
G	geographical names of any kind
m	product, organization, company and other proper names
U	medical, chemistry and natural science terms

Table 3: List of semantic labels.

3.5 Derivative Relations

The word-formation relations in Czech has been delegated to derivational data sources, such as Derinet (Vidra et al. 2019).[6] In MorfFlex, the lemma contains information about the base lemma it is derived from only in case of regular derivations. For example, lemmas of possessive adjectives (e.g. lemma: *otcův_^(*3ec)*) contain information about the noun they are derived from: *otcův 'father's' → otec 'father'*. The originating lemma is (for space saving reasons only) written in the form of a rule. For example, derivation information **3ec* in lemma *otcův_^(*3ec)* means remove 3 characters, add *ec* to get *otec*.

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

4 Analysis: From Wordform to its Detailed Morphological Description

For each wordform, the structured morphological information is captured in its tag. Each position of the tag captures a different aspect of the wordform, e.g. its case or number, the style of the inflectional variant, or a characteristic of the lemma equivalent to information given in typical monolingual dictionaries, e.g. part of speech, detailed features of the particular part of speech (such as possessivity of pronouns, verbal aspect, animateness of nouns etc.).

4.1 Special Parts of Speech

The values on the individual tag positions reflect the morphological focus of the dictionary. Therefore, in addition to the standard POS set, several special categories are used (cf. Hlaváčová et al. 2019; Mikulová et al. 2020): foreign word, segment, abbreviation, and isolated letter. The new categories of POS allow to describe the diversity of the language much more precisely.

4.1.1 Foreign word

Foreign word (F at the POS tag position) identifies a word that is not subject to the Czech inflectional system, often creates a part of a longer foreign phrase in a Czech text and has no meaning of its own in Czech.

4.1.2 Segment

Segment (S at the POS tag position) describes a part of a word that creates a complete meaningful unit only when joined with another component. In MorfFlex, we distinguish two types of segments. First, the so called *prefixal segments,* which stand at the beginning of a word and which express no morphological categories. Secondly, the so called *postfixal segments,* which may express all morphological categories of a particular part of speech. Table 4 shows the analysis of the segments of the tokenized compound adjective *tchaj-pejský* (*'Taipei['s]'*). The segment *pejský* behaves as an adjective and expresses the morphological features of the compound adjective.

Wordform	Lemma	Tag
tchaj	tchaj	S2--------A----
pejský	pejský_^(tchaj-pejský)	SAMS1----1A----

Table 4: Segments.

4.1.3 Abbreviation

Abbreviations (B at the POS tag position) composed of capital letters and representing a multi-word unit (e.g. *ČR* for *Česká republika 'the Czech Republic'*) are considered to be separate parts of speech. In contrast, the abbreviations abbreviating a one-word term are captured as a special wordform in the paradigm of the term, i.e., by the letter *b* at the 15th position of the tag (e.g. *čt.* for *čtvrtek 'Thursday'*). See examples in Table 5.

Wordform	Lemma	Tag
USA	USA_;G_^(United_States_of_America)	BNXXX-----A----
ČT	ČT_;m_^(Česká_televize)	BNXXX-----A----
čt	čtvrtek	NNIXX-----A---b
m	minuta	NNFXX-----A---b

Table 5: Abbreviations.

4.1.4 Isolated letter

Isolated letters (Q at the POS tag position) stand for many meanings but it is not clear for which of the many alternatives. We do not distinguish between an abbreviation (e.g. *A. Franklin*) and a label (e.g. *skupina A 'group A'*, *A-konto 'A-account'*) and between the other meanings such as for sorting a list (e.g. *a), b)*) or as a graphical separator in a text (e.g. *o o o o o o o o*). See examples in Table 6. The introduction of new POS for isolated letters does not mean that standard POS such as conjunctions or prepositions are not distinguished for one-letter word (e.g. letter *a* in *otec a matka 'father and mother'* is considered a conjunction POS).

Wordform	Lemma	Tag
(skupina) A	A-33	Q3----------
A. (Franklin)	A-33	Q3----------

Table 6: Isolated letters.

4.2 Homonymy of Wordforms

Searching by wordform also helps in the recognition and interpretation of homonymous forms, both within the lemma and across the whole dictionary. As it is evident from Table 7, the wordform *sil* is analyzed as the genitive plural form of two different nouns, *silo 'silo'* and *síla 'power'*, as the masculine singular past participle of the verb *sít 'to plant'*, and as two forms of the imperative of the verb *sílit 'to strengthen'*.

Wordform	Lemma	Tag
sil	sít_^(zasévat_[semena,...])	VpYS----R-AAI--
sil	silo_^(pro_úschovu_např._krmiva;_raket)	NNNP2-----A----
sil	síla_^(fyzická,_vojenská;_moc)	NNFP2-----A----
sil	sílit_^(získávat_sílu)	Vi-S---2--A-I--
sil	sílit_^(získávat_sílu)	Vi-S---3--A-I-4

Table 7: Homonymy of wordforms.

5 Conclusion

We have demonstrated the possibilities of the exploitation of the MorfFlex dictionary for linguistic research purposes. In contrast to other Czech monolingual dictionaries or grammar handbooks, MorfFlex contains not only much more morphological data, described in detail (as expected), but it also covers a wider range of words, including non-standard wordforms and lemmas. Furthermore, the dictionary is extended by adding semantic labels and stylistic labels and other complementary tools, which firstly serve to specify and clarify morphological data, and secondly to simplify the orientation in the dictionary for users. Although the dictionary also provides some wordforms that are only potential (i.e. unattested), due to a significant proportion of manual annotation and consequent unification the results are relatively reliable, therefore the dictionary could serve as a resource for diverse linguistic research, and as a morphological support for the creation of other dictionaries.

6 References

Benko, V. (2014). Aranea: Yet Another Family of (Comparable) Web Corpora. In Sojka, P., Horák, A., Kopeček, I., Pala, K. (eds): *TSD 2014*, LNAI 8655. Springer International Publishing, pp. 257–264.

Bermel, N. (2000). *Register Variation and Language Standards in Czech*. Studies in Slavic Lingustics, 13. Muenchen: LINCOM EUROPA.

Bermel, N. (2007). *Linguistic authority*, *language ideology, and metaphor: the Czech orthography wars*. Berlin-New York: Mouton de Gruyter.

Ferguson, Ch. A. (1959). Diglossia. *Word,* 15, pp. 325-340.

Hajič, J. (2004). *Disambiguation of Rich Inflection*. *(Computational Morphology of Czech).* Prague: Karolinum.

Hajič, J., Hajičová, E., Mikulová, M., Mírovský, J. (2017). Prague Dependency Treebank. In *Handbook on Linguistic Annotation.* Dordrecht*:* SpringerVerlag, pp. 555–594.

Hajič, J., Hlaváčová, J., Mikulová, M., Straka, M., Štěpánková, B. (2020a, in press). *MorfFlex CZ.* Institute of Formal and Applied Linguistics, LINDAT/CLARIAH-CZ, Charles University, Prague, Czech Republic, LINDAT/CLARIAH-CZ PID: http://hdl.handle.net/11234/1-3186.

Hajič, J., Bejček, E., Hlaváčová, J., Mikulová, M., Straka, M., Štěpánek, J., Štěpánková, B. (2020b). Prague Dependency Treebank - Consolidated 1.0. In *Proceedings of The 12th Language Resources and Evaluation Conference*, European Language Resources Association, Marseille, France, pp. 5208—5218.

Hlaváčová, J., Mikulová, M., Štěpánková, B., Hajič, J. (2019). Modifications of the Czech morphological dictionary for consistent corpus annotation. *Jazykovedný časopis/Journal of Linguistics,* 70 (2), pp. 380–389.

Hoffmannová, J. (2013). Česká hovorovost a hovorová čeština (v kontextu dalších slovanských jazyků). *Slavia* 82, pp. 125–136.

Kosem, I., Lew, R., Müller-Spitzer, C., Ribeiro Silveira, M., Wolfer, S., Dorn, A. et al. (2019). The image of the

monolingual dictionary across Europe. Results of the European survey of dictionary use and culture. *International Journal of Lexicography*, 32 (1), pp. 92-114.

Mikulová, M., Hlaváčová, J., Hajič, J., Hana, J., Hanová, H., Hladká, B., Štěpánková, B., Zeman, D. (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. In press.

Skoumalová, H., Hnátková, M., Petkevič, V. (2011). Linguistic Annotation of Corpora in the Czech National Corpus. In Zacharov, V.: *Trudy meždunarodnoj konferencii "Korpusnaja lingvistika – 2011"* (*Proceedings of the International Conference "Corpus Linguistics – 2011")*. St.-Peters burg State University, Institute of Linguistic Studies, Sankt-Petěrburg, Russian State Herzen Pedagogica, pp. 15-20.

Slovník spisovné češtiny. (1978) (Second, revised edition 1994; third, revised edition 2003). Prague: Academia.

Slovník spisovného jazyka českého. (1960–1971). Prague: Academia.

Straková, J., Straka, M., Hajič, J. (2014). Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition. In *Proceedings of 52nd Annual Meeting of the ACL: System Demonstrations*. Association for Computational Linguistics, Baltimore, pp. 13-18.

Šipková, M. (2017). Interdialekt. In P. Karlík, M. Nekula, J. Pleskalová (eds.), *CzechEncy - Nový encyklopedický slovník češtiny*. https://www.czechency.org/slovnik/INTERDIALEKT [20/05/2020].

Vidra, J., Žabokrtský, Z., Ševčíková, M., Kyjánek, L. (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*. Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics, Prague, Czech Republic, pp. 81–89.

Acknowledgements

The research and language resource work reported in the paper has been supported by the LINDAT/CLARIAH-CZ projects funded by Ministry of Education, Youth and Sports of the Czech Republic (project LM2018101).