sciendo
DOI 10.2478/jazcas-2023-0045

LEMMATIZATION OF THE DIA1900 DIACHRONIC CORPUS

LUCIE BENEŠOVÁ[1] – KLÁRA PIVOŇKOVÁ[1,2]
– MARTIN STLUKA[1]

[1] Institute of the Czech National Corpus, Faculty of Arts, Charles University, Prague, Czech Republic
[2] Department of Philosophy and History of Science, Faculty of Science, Charles University, Prague, Czech Republic

BENEŠOVÁ, Lucie – PIVOŇKOVÁ, Klára – STLUKA, Martin : Lemmatization of the DIA1900 Diachronic Corpus. Journal of Linguistics, 2023, Vol. 74, No 1, pp. 275 – 284.

Abstract: This paper focuses on the process of lemmatization of the upcoming Czech diachronic corpus of the second half of the 19th century, DIA1900. The article describes different approaches to the corpus lemmatization of synchronic written, spoken and diachronic corpora within the Czech National Corpus project, including single- and multilevel lemmatization and available tools used to link the variants.

Keywords: Czech, diachronic corpus, disambiguation, lemmatization, morphological dictionary, variability

1 INTRODUCTION

The Czech National Corpus[1] (CNC) offers free access to corpora of various types. These include not only large synchronic corpora of written language (SYN series), corpora of spoken language (ORAL, ORTOFON, DIALEKT), but also various types of specialized corpora and the diachronic corpus DIAKORP. The aim of this paper is to describe the concept of lemmatization of another upcoming diachronic corpus, DIA1900.

2 CORPUS DIA1900 AND ITS SPECIFICS

The DIA1900 corpus comprises 103 full texts from the second half of the 19th century and contains a total of approx. 1 million word forms.[2] It is designed to be as representative as possible. It is genre-balanced (fiction/non-fiction/newspaper and magazines) in each decade, including consideration of the balance of length and number of texts. The corpus will be fully lemmatized and morphologically annotated.

[1] https://korpus.cz/
[2] At present (March 2023), this corpus is in its penultimate (disambiguation) phase.

Its aim is to faithfully capture historically relevant linguistic variability, and at the same time in such a way that variation will present as little of an obstacle to the user as possible. The second half of the 19th century represents a specific period in the development of the Czech language, involving several distinctive linguistic phenomena. The gradual expansion of the functional range of Czech in many disciplines was accompanied by the rise of linguistic variability (Kučera et al. 2019). This includes word forms that, although understandable from a contemporary perspective, are unpredictable. For these reasons, the decision was made not to use procedures and tools common in the creation of synchronic corpora when building the DIA1900 corpus (see Benešová et al. 2023).

3 MULTILEVEL LEMMATIZATION AND CORPORA OF DIFFERENT TYPES

Lemmatization, a part of the process of morphological annotation, consists in the assignment of a lemma to one word form. Before we focus on the actual description of the lemmatization of the DIA1900 corpus, it is necessary to mention the change introduced in the field of lemmatization by the latest corpora of the SYN series. Starting with the SYN2020 corpus, lemmatization in the CNC is two-level. In addition to the standard lemma, each form is assigned a sublemma attribute.[3] While a lemma can cover more variants of a single word (e.g. the lemma *filozofie* 'philosophy' represents both forms with *-z-* (*filozofie*) and with *-s-* (*filosofie*)), sublemmas define subsets of forms according to this variation (the sublemma *filozofie* represents only forms with *-z-* (*filozofie*), the sublemma *filosofie* only forms with *-s-* (*filosofie*) in all cases).

Different types of invariance are handled by these sublemmas in synchronic written corpora: e.g. orthographic variation; spelling variation and alternations (*citron*/*citrón* 'lemon') and variants of other types; irregular forms and special categories of gradations (*dobře/lépe/líp* 'well/better'); nominal forms of adjectives and others.

Multilevel lemmatization is an excellent option for solving a number of linguistic dilemmas related to the creation of corpora of different types. The lemmatization of diachronic data, as well as e.g. spoken data processing,[4] is far more challenging than in the case of written synchronic language due to their greater linguistic variability.

The linguistic material of spoken and diachronic corpora could be treated using the aforementioned multilevel lemmatization, and this would be a viable solution. However, at the time the lemmatization concept for the DIA1900 corpus was developing, a two-level lemmatization as implemented in e.g. the SYN2020 corpus

[3] For details, cf. https://wiki.korpus.cz/doku.php/cnk:syn2020:lemmatizace.

[4] The concept of lemma is adapted in these corpora so that all forms realized in diverse graphical and sound variants can be found by the user. Within the spoken language corpora, for example, the lemma *tenhleten* 'this' includes a total of 105 word forms; in Nsg. neuter the following forms occur: *tohleto, todnecto, todleto*, *todlecto*, *todlencto*, *tohlencto*, *tohlento*, *toleto*, *tohlensto*, *todlensto*.

(cf. Hlaváčová et al. 2019) was not yet an option.[5] In the context of diachronic corpora, the need to interlink variant sets of paradigms was even stronger than in synchronic corpora, given that it is a solution that can make historically symptomatic variants available to the user. It was necessary to develop a concept that would, even with a single-level lemmatization, provide a versatile model of variant mapping that also takes into account the typology of the linguistic phenomena reflected in the variants.

4 VARIATION AND FORM OF THE LEMMA IN THE DIA1900 CORPUS

4.1 Morphological dictionary

A morphological dictionary (MD) is a collection of individual word forms with lemmas and morphological tags. A synchronous morphological dictionary is used in the automatic creation process of synchronic corpora.

Although we are aware of the alternative approaches to compiling diachronic corpora[6] (for example, with regard to the distribution of automatic processes used in text preparation and lemmatization), the concept of the DIA1900 corpus reckoned with the completely new creation and generation of a large custom-made morphological dictionary. The resulting MD contains 23.5 million word forms, and the lexical material originates from various sources including 19th-century dictionaries and texts. The tag basis of this MD is the Prague (Hajič's[7]) tagset. The original tagset was modified in several positions for the specific purposes of the DIA1900 corpus (a detailed description of the morphological dictionary is beyond the scope of this paper – for more details cf. Benešová et al. in 2023). This allowed the use of lemmatization not only for the simple preservation of period variability, but also for linking variants with synchronic standard lemmas so that even forms unpredictable to contemporary speakers could be made available. The linking of variants concentrated mainly on phenomena that are not related to word formation and that are not included in the contemporary dictionary.[8]

The MD was tailored to the chosen concept of lemmatization. The creation of the MD was motivated by a desire to capture the wide variability of 19th-century language, which could not be adequately treated by any of the existing synchronic tools, i.e., without a large error rate.

---

[5] The concept of two-level lemmatization of the SYN series corpora was developed later and independently of the concept of lemmatization of the DIA1900 corpus.

[6] Cf. e.g. Hana et al. 2012; Kieraś et al 2018.

[7] Cf. Hajič 2004.

[8] It was not the aim in this respect to cross the word-formation and lexical boundary; lemmas such as *propůjčník,* 'lender', *zakazovatel* 'forbidder' or *vodychudý* 'water-poor' (all obsolete) will not be linked in any way to potential equivalents that are part of synchronic language usage, since such a link would be unreliable and rather confusing without a detailed lexicographical analysis.

Variation is captured in the MD both at the level of forms and at the level of lemmas.[9] The form-level variation is characterized by the existence of overabundance (cf. Thornton 2019) within the same paradigm (the synchronic aspect of standard or non-standard variation is not relevant). A typical case of form-level variability is e.g. the form of the word *tráva* 'grass' for INS sg. (*trávou*/ *travou*).[10]

Lemma-level variability refers to all word forms related to a single paradigm. The problem of how to treat the variability of this type is ever-present in diachronic corpora due to the spread of linguistic changes over time. The MD framework reflects 2 types of this variability, which are implemented in the corpus by the attribute of lemma, respectively hyperlemma[11] and the Search Suggestions Tool (hereafter SST) (see Section 4.2 for details).

The MD also includes the generated morphological categories in a different form and scope in comparison with contemporary Czech. For example, for the past and present transgressive it is necessary to assume the 19th-century interpretation that each of the forms of these transgressives is capable of expressing any gender and number.[12] Such a broader conception of certain morphological categories allows us to decide in favour of function over form in specific cases (e.g. within a morphological category, the form of the transgressive *dělaje* 'doing' can be feminine in the sg.).

The MD enabled us to identify linguistic phenomena typical of an earlier phase of Czech, and the modified tagset allowed us to mark categories evaluated as relevant and worthy of specification.[13] Providing a detailed description of all the tagset specifics of the DIA1900 corpus goes beyond the scope of this paper, but at least two examples are worth mentioning: marking the proprial character in the 6th position and marking the affixed particle in the 15th position of the tagset.[14]

The success rate of MD when first applied to the corpus text material was 97%, and further modifications increased this score.

---

[9] Within the context of the cited paper by Hlaváčová et al. (2019), we can refer to inflectional and paradigmatic variability.

[10] Quantity is the most frequent source of variation in the 19th century, both at the level of forms and at the level of lemmas.

[11] On the concept of hyperlemma, cf. https://wiki.korpus.cz/doku.php/pojmy:lemma; e.g. the forms *bejt* and *být* have the lemma *být* 'be' assigned throughout their paradigms.

[12] The usage of past and present transgressive was considerably divergent in the 19th century due to conflicting tendencies, one of which led to grammaticalization in the language and the other to their rehabilitation in the state that existed about two centuries earlier.

[13] For a detailed description of the tagset values of the DIA1900 corpus specifics cf. Benešová et al. (2023).

[14] The usage of the affixed particle -ť, -tě, -(i)ž is a frequent linguistic phenomenon in the 19th century (***Bylotě** to jeho přání* 'it **was** his wish'). This has been reflected in the MD – the forms with the affixed particles have been generated to all verbal and adjective paradigms, e.g. *bylotě být* VpNS------AA--**T**I.

4.2 Linking variants

In order to link variants and to facilitate the search, three procedures were reserved in the lemmatization process of the DIA1900 corpus:

1. Assigning an individual form, set of forms, or the whole paradigm to a synchronically standard lemma. One lemma has more than one paradigm in the MD in the following cases:

(a) Regular vowel variations of a vocalic nature (*ú>ou, ý>ej*, prothetic *v* and, to a limited extent, *é>í*) are handled in this way – these variations are assigned to the lemma that represents the synchronic standardized variant. Thanks to this procedure, they are both preserved and traceable.

(b) This procedure is also used in cases which, within the context of linguistic and lexicographic convention, appeared to be a user-friendly solution: comparative and superlative forms for adjectives and adverbs are assigned to the positive, nominal forms of adjectives to the standard adjectival lemma, and the negative paradigm for verbs is assigned to the affirmative lemma[15] (see Tab. 1, Section 7).

2. The second approach is to draw the user's attention to the relationship between the lemmas via the search engine user interface, more specifically by the SST. This tool offers variant lemmas during the process of setting a corpus query. Typically, these are variants in the domain of spelling, or for example variants resulting from consonant alternation and phenomena related to variations in vocalic quantity,[16] manifested in the whole paradigm[17] (see Tab. 1, Section 5).

3. The third technique used is the assignment of more than one lemma to a single form. These are both cases of aggregates that contain more lemmas in a single token (*uvěřils* > *uvěřit_být* 'you believed > to believe_to be') and so-called "multi-lemma", i.e., capturing form-level homonymy in semantically equivalent variants where the lemma cannot be determined from the context. Although this procedure is not directly related to the user's query input, it provides the user with information about the existence of variability that might be unintuitive to the contemporary native speaker (see Tab. 1, Sections 2 and 3; see more detail in Section 5).

---

[15] The area of negation signalled by the prefix of *ne-* 'not' was generally treated as follows in the lemmatization: negative verb forms were assigned to the affirmative lemma and have a marked negation in the 11th position of the tag. For adjectives, the forms with negation were given the negative lemma and have negation marked in the 11th position of the tag. Nouns beginning in the prefix *ne-* also have a lemma in *ne-*, but due to more advanced lexicalization, negation is no longer marked in the 11th position of the tag.

[16] Lemmas that are non-standard from today's point of view, and hence harder to predict, will be offered during query entry even if the corpus does not contain a synchronically standard lemma. E.g. the DIA1900 corpus does not contain the *užívatel* 'user' lemma, but we do find here the forms of the *užívatel* 'user' lemma. The existence of a lemma with an irregular quantity is pointed out to the SST when the synchronously standard variant *uživatel* 'user' is entered.

[17] Where the deviation is restricted to a single form (e.g. for alternation in the durative participle of verbs, for quantity in single forms in paradigms where the alternation of quantity is standard), the variant form is assigned as a doublet (procedure 1.b, see Tab. 1, Section 6).

LIST OF EXAMPLES			
	Token	**Lemma**	**SST**
I. *regular changes of vocalic character*	*kdož jste vy podlého* ***smejšlení*** 'who are you of a mean **mind**'	smýšlení 'mind, opinion'	no
	smýšlení *o tom rozdílné panuje* 'there is a difference of **opinion** about it'		
	nalézáme jména: [...] štika, kapr, ***vokoun*** 'we find the names: pike, carp, **perch**'	okoun 'perch'	no
	nelze rozeznati, je-li ***okoun*** *mlíčný* 'it is not possible to distinguish if the **perch** has a milt'		
	nepřítele ***nejouhlavnějšího*** *pouštíš tak na svobodu* 'you're letting your **main** enemy go free'	úhlavní 'main'	no
	na [...] Pinovi spočívalo podezření co ***úhlavním*** *spachateli* 'Pino was suspected of being the **main** perpetrator'		
II. *form-level homonymy for lemmas that are synonymous and predictable to the contemporary speaker*	*Černá* ***země*** *ssaje krve potoky* 'the black **earth** sucks streams of blood'	země 'earth'	no
	pálená ***zem*** *s popelem se roztrousila* 'the scorched **earth** with ashes was scattered'	zem 'earth'	no
	zasype je ***zemí*** 'he'll cover them with **earth**'	země/zem 'earth'	no
III. *form-level homonymy for lemmas that are synonymous and include a variant that is potentially unpredictable to the contemporary speaker*	*umrtvuje v sobě ruch* ***duchovný*** 'he mortifies **spiritual** activity in itself'	duchovný 'spiritual, clerical'	yes
	duchovní *ruch poměrně živěji v Evropě [se] osvědčoval* '**spiritual** activity proved to be relatively more lively in Europe'	duchovní 'spiritual, clerical'	yes
	duchovní *hodnostáři* '**clerical** dignitaries'	duchovný/duchovní 'spiritual, clerical'	yes
IV. *phenomena covering equivalent variants (typically spelling) that are potentially unpredictable to contemporary native speakers*	*Tento* ***způsob*** *sázení* 'this **way** of planting'	způsob 'way'	yes
	dosavadní ***způsob*** *panování* 'the existing **way** of ruling'	způsob 'way'	yes
	Černá péra [...] můžeme dle ***methody*** *této úplně vybíliti* 'black feathers can be completely bleached out using this **method**'	methoda 'method'	yes
	Návodův a ***metod*** *ke cvičením zpěvu zná každý* 'everyone knows the instructions and **methods** to practice singing'	metoda 'method'	yes

LIST OF EXAMPLES			
	Token	**Lemma**	**SST**
V. *phenomena related to e.g. vocalic quantity or consonant alternation – the whole paradigm*	*když dosavádní **užívatel** dobrovolně zákazu [...] se podrobí* 'when an existing **user** voluntarily submits to a ban'	užívatel 'user'	yes
	*Mezi Rakouskem a Bavory je uzavřeno **usnešení*** 'a **resolution** is concluded between Austria and Bavaria'	usnešení 'resolution'	yes
	*aby se **usnesení** o peticích odložilo* 'that the **resolution** on the petitions be postponed'	usnesení 'resolution'	yes
VI. *phenomena related to e.g. consonant alternation or quantity – single form*	*Bylť jest [...] ze země **vypovězen*** 'he was **expelled** from the country'	vypovědět 'expel'	no
	*ze země **vypověděna** byla* 'she was **expelled** from the country'		
	*Měřice **hráchu*** 'scoops of **peas**'	hrách 'peas'	no
VII. *negation, gradation, nominal adjectives*	*nutili otroky své ku […] **nepravostem*** 'they forced their slaves into **iniquity**'	nepravost 'iniquity'	no
	*vše připraveno bylo rukou [...] **nešťastných** spolubratrů* 'it was all set up by the **unfortunate** fellows'	nešťastný 'unfortunate'	no
	*vina **nepochybně** padá na manželku* 'the blame **undoubtedly** falls on the wife'	nepochybně 'undoubtedly'	no
	*Ploutev […] **má** načervenalý kraj* 'the fin **has** a reddish edge'	mít 'to have'	no
	*jednotlivý víc **nemá** místa* 'the individual **has no** place anymore'		
	*krmiti ji musíme hojně a **dobře*** 'we have to feed her **well** and abundantly'	dobře 'well'	no
	*jeho zboží [...] **nejlépe** se doporučuje* 'his goods are **best** recommended'		
	*Zůstals **věren** učení* 'you have remained **faithful** to the teachings'	věrný 'faithful'	no
	*Vítej mi, **věrný** strážce* 'welcome, **faithful** guardian'		

Tab. 1. Methods of capturing variability in the DIA1900 corpus

5 DISAMBIGUATION

After the MD was applied to the texts, the next step was the creation of an etalon corpus with 500 thousand tokens (the size of the etalon corpus corresponds to the needs of later automatic annotation, as well as to the personnel and time constraints).[18] The process of lemmatization and manual disambiguation consists not only in the selection of a single-valued lemma for a given form, but often also in deciding on the assignment of a multi-lemma.

5.1 Single-lemma choice

Concerning the ambiguous forms, we distinguish between homonymy that can be disambiguated by the context and homonymy where our decision can only be based on knowledge of the language system of a given time period and is necessarily inherently associated with uncertainty. In other words, we are able to resolve the homonymy when we have sufficient relevant evidence to make a decision, thanks to a sufficiently comprehensive context and semantics. Based on the context that identifies the meaning of a given token, we are able to assign an adequate lemma and tag, e.g. in the context *valně se liší od* ***divákův*** 'largely differs from **spectators**' the form *divákův* is the noun *divák* 'spectator', not the possessive adjective *divákův* 'spectator's'.

The problem of homonymy arises when the language system involves equivalent variants within the same paradigm and the context cannot be used to specify them. At the same time, it is necessary to comply with the requirement that all categories that are part of a tag are unambiguously specified. As a result, a decision in favour of one of the variants is necessary. Therefore, specific disambiguation rules regarding these contentious solutions have been developed, taking into account the assumed language and grammar of the 19th century.[19] Even so, in many cases it is not possible to make a clear-cut decision with certainty. This is the case, for example, with the case homonymy of the genitive and accusative (*nesnadno si představiti* ***člověka*** *tak lhostejného* 'it is not easy to imagine a **man** so indifferent').[20]

5.2 Multi-lemma choice

However, the specific rules adopted, together with the technical solution of lemmatization, also allow us to assign multi-lemmas (each with their own tag) to a single token. As an example, we can use the form *křeče* 'cramp, spasm' in the sentence *stonala 15 let na žaludkové* ***křeče*** 'she suffered from stomach **cramps** for 15 years', where we have the possibility of assigning both a lemma and a tag for the

[18] The application of the MD resulted in 43.45% of unambiguous tokens and 56.55% of ambiguous tokens that need to be manually disambiguated.

[19] E.g. Pravidla (1902) and (1913), and Jungmann (1835–1839), PSJČ (1935–1957).

[20] In 19th-century Czech, there are some linguistic phenomena which are no longer known in contemporary Czech and which do not make the decision easier, e.g. the genitive of negation.

feminine and the masculine to a given token, i.e., *křeč* ‘cramp’ N-FP4-----A----- and *křeč* ‘cramp’ N-IP4-----A-----). However, assigning multiple lemmas to a single token also has limitations. If we disambiguate the form of the adjective ***žaludkové*** ‘stomach’ in the above sentence, we need to choose either the masculine or the feminine (due to the adjectival character of the lemma, the forms of both genders belong to the same lemma). In these cases, we choose the gender variant of the adjective that is more prevalent in the text with the controlling noun in unambiguous contexts. If we do not have contextual support, we take into account the linguistic usage of the 19th century or the contemporary usage.

6 CONCLUSION

Building a diachronic corpus is always a complex challenge. The DIA1900 corpus represents one possible solution to such a task. In principle, the concept of lemmatization of the DIA1900 corpus offers the user access to historically characteristic variability, either through directly assigned synchronously standard lemmas or through the SST when entering a corpus query. Thanks to the chosen measures in the processing of the DIA1900, in particular the creation of a manually disambiguated etalon that can be used as training data for the tagger (automatic annotation), a solution has been provided that will enable easier automatic processing of other corpora in future, even using multilevel lemmatization.

ACKNOWLEDGEMENTS

This paper resulted from the implementation of the Czech National Corpus project (LM2023044) funded by the Ministry of Education, Youth and Sports of the Czech Republic within the framework of Large Research, Development and Innovation Infrastructures.

References

Benešová, L., Kučera, K., Najbrtová, K., Pivoňková, K., and Stluka, M. (2023). Korpus DIA1900: jeho koncepce a vytváření. Časopis pro moderní filologii, 105, pages 121–140.

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

Hana, J., Lehečka, B., Feldmann, A., Černá, A., and Oliva, K. (2011). Building a Corpus of Old Czech. In Proceedings of the Adaptation of Language Resources and Tools for Processing Cultural Heritage Objects Workshop associated with the 8th International Conference on Language Resources and Evaluation (LREC 2012), pages 9–15.

Hlaváčová, J., Mikulová, M., Štěpánková, B., and Hajič, J. (2019). Modifications of the Czech morphologigal dictionary for consistent corpus annotation. Journal of Linguistics, 70(2), pages 380–389.

Jungmann, J. (1835–1839). Slovník česko-německý. Praha. Accessible at: https://vokabular.ujc.cas.cz/moduly/slovniky/.

Kieraś, W., and Woliński, M. (2018). Manually Annotated Corpus of Polish Texts Published between 1830 and 1918. In Proceedings of the Eleventh International Conference on Language Recources and Evaluation (LREC 2018). Miyazaki, Japan: ELRA, pages 3854–3859.

lemma. https://wiki.korpus.cz/doku.php/pojmy:lemma (Last visited March 24, 2023).

lemmatizace. https://wiki.korpus.cz/doku.php/cnk:syn2020:lemmatizace (Last visited March 24, 2023)

Kučera, K., Najbrtová, K., Pivoňková, K., Řehořková, A., and Stluka, M. (2019). Korpus českého jazyka 2. poloviny 19. století. Časopis pro moderní filologii, 101, pages 92–97.

Příruční slovník jazyka českého (PSJČ) (1935–1957). Accessible at: https://bara.ujc.cas.cz/psjc/.

Pravidla hledící k českému pravopisu a tvarosloví s abecedním seznamem slov a tvarů (1902). Praha: Školní knihosklad.

Pravidla českého pravopisu s abecedním seznamem slov a tvarů: Jediné c.k. ministerstvem kultu a vyučování schválené vydání (1913). Praha: Školní knihosklad.

Thornton, A. M. (2019). Overabundance in Morphology. In Oxford Research Encyclopedia of Linguistics. Oxford University Press.