The Sloleks Morphological Lexicon and its Future Development

Kaja Dobrovoljc, Simon Krek and Tomaž Erjavec

Abstract

This paper presents Sloleks, the largest open-source machine-readable morphological lexicon of the Slovene language to date. We first briefly present its development and the formal grammar behind it, and then provide a detailed presentation of the types and structure of inflectional, derivational, grammatical and other included information, with a special emphasis on its formal representation within the standardized XML LMF framework. Given that Sloleks is a strong candidate to be used in the compilation of a new dictionary of modern Slovene, both as a source of morphological information and as a background resource for the language technology tools needed to create it, the second part of the presentation explores the most important aspects of its future development, in particular the expansion of its entry list, addition of pronunciation information, normative categorization of variants and a corpus-based re-evaluation of the existing inflectional paradigms. Such an extensive usage-based open-source morphological lexicon of modern Slovene with a unified system of morphological description will have a long-term use for both language technologies and for other born-digital reference works for the Slovene language.

Keywords: morphological lexicon, lexicon of inflected forms, machine readable dictionary, morphology, inflection, derivation, pronunciation, language standardisation

1 INTRODUCTION

When it comes to morphologically rich languages, such as Slovene, the description of morphological paradigms of inflected parts of speech is traditionally very important. For example, the first Slovenian grammar (Bohorič 1584) dedicates almost half of its content to word inflection, and morphological paradigms have a similarly prominent role in most of the later Slovene grammars. These mainly focus on systemic aspects of morphology, i.e. morphological patterns which they illustrate by means of examples. This in turn means that explicit, complete paradigms in grammar books are few and far between. On the other hand, dictionaries from the pre-digital age, mainly different orthography guides and later DSLL (*The Dictionary of Slovene Literary Language*), fulfilling their role as lexical enumerators, also contained data on inflection. The morphological descriptions in these reference works are significantly shortened; in addition to the headword, they are usually limited to one or a few inflectional forms, which are supposed to provide the user with enough information to deduce the entire morphological paradigm. Even when printed reference books were digitized, the data stayed the same.

The arrival of computers and advances in natural language processing soon established a need for accessible machine-readable dictionaries and lexicons of inflected forms (Atkins and Zampolli 1994). The first English machine-readable dictionaries designed for various language technology tasks were already designed in the 1960s (e.g. Boguraev and Briscoe 1987); the widespread digitization of languages in the 1990s, however, also paved the way for the creation of morphological lexicons for most other European languages.

Computers cannot work with only a pattern or a few word forms, which is why these lexicons – free from the space constraints imposed by the printed medium – typically contain paradigms written out in full and available in a machine-readable format. Morphological data, traditionally targeted at users of printed language reference books, were therefore given a new field of application, where the new "user" is the computer itself. Lexicons must therefore fulfil language technology needs in various computer applications – from spellcheckers and part-of-speech taggers to parsers, speech synthesizers, and machine translation software – and be simultaneously useful as independent morphological reference tools for language users. The contemporary machine-readable lexicon of the Slovene language should therefore fulfil both needs, and thus needs to be organized differently than morphological data in dictionaries and grammar books or the first computational lexicons.

In pursuing these two goals, the compilation of such lexicons stumbles upon two contradictory tendencies: when dealing with language technology applications,

the lexicon must be capable of representing the morphological characteristics of all the word forms present in authentic texts, including spoken discourse, allowing for simple machine processing of the data. However, when it comes to traditional usage, it must also provide effective information on inflection, pronunciation, and word derivation relevant for a human user, including normative aspects of the vocabulary. In the context of integrating a lexicon into a future dictionary of modern Slovene, the lexicon's content must be aligned with both poles: on the one hand with the morphological data produced by morphological taggers to automatically annotate text corpora (the data source of the dictionary), while also making sure that the lexicon aligns with the data in the lexical database used as the source of the dictionary.

When it comes to fulfilling the user needs associated with language reference books, the key problem in creating the reference morphological lexicon of modern Slovene lies in the fact that the existing language reference grammar books (e.g. Toporišič 2004), dictionaries (e.g. DSLL2) and normative guides (e.g. SP 2001) are not on the same page when it comes to examining morphological data, and at times even contradict one another (cf. Krek 2014a). This means that none of this work can be taken as a starting point – the whole concept needs to be redesigned from scratch. Additionally, these reference books were not created based on modern language data, meaning they are relatively detached from the linguistic reality of modern Slovene, although this is important for users of language reference books and for language technologies.

Computational morphological lexicons for Slovene have a relatively long history. At the start of the 1990s, the Amebis company started developing ASES, an electronic dictionary of the Slovene language, which also contains explicit morphological paradigms (Arhar and Holozan 2009). This database itself is not freely available; however, the data it contains may be found in various products the company offers, such as the Besana grammar checker, the Presis machine translation software, its system for natural language communication, and so on. Chronologically speaking, the first freely accessible computational lexicon of Slovene was created in the framework of the MULTEXT-East project in the 1990s. It contains over 15,000 lemmas and their inflectional paradigms in a tabular format (Erjavec et al. 1995).

During the first decade of this century, the development of speech technology (mainly speech synthesis) raised the importance of lexicons which – in addition to morphological data – also contain information on pronunciation, such as SIflex, SImlex (Rojc et al. 2002; Verdonik et al. 2002), LC-STAR (Verdonik et al. 2004; Verdonik and Rojc 2004), SI-PRON (Žganec Gros et al. 2006). The chief problem with all these lexicons lies in the fact that they are not freely available. The same goes for the morphological lexicon created during the same period at

the Fran Ramovš Institute of the Slovenian Language. There are in fact no data available on this lexicon, apart from the fact it exists (Naglič et al. 2005: 36).

A slightly more specific lexicon is available through the freely accessible machine translation system called Apertium; it contains just over 20,000 lemmas (Horvat and Vičič 2012; Vičič 2012). Even though it is basically derived from the MULTEXT-East lexicon, its content and format is somewhat different, since it is mainly used in the context of a translation system, and is therefore not useful as a general morphological lexicon for Slovene. Within the recently completed "Communication in Slovene" project, the morphological lexicon Sloleks (Dobrovoljc et al. 2013) was created. This is also the central subject of this chapter – because due to its size, accessibility, and use in Slovene language technology tools, it represents a logical stepping stone for the further development of a reference morphological lexicon for Slovene.

2 THE SLOLEKS MORPHOLOGICAL LEXICON

The following sections describe the content of the Sloleks morphological lexicon and its format, the types of data it contains and their organisation within an individual lexicon entry, and the design of its online interface.

2.1 Content

2.1.1 Lemma list and paradigms

The current version of Sloleks (Dobrovoljc et al. 2013) includes 100,805 entries, where an entry includes the basic form (the lemma) of the word, its inflected forms (the inflectional paradigm) and related morphological information. The list of headwords or lemmas has been compiled based on criteria set out in the guidelines for its construction (Erjavec et al. 2008), by first including the majority of lemmas occurring in the manually annotated ssj500k corpus (Krek et al. 2013c), all lemmas belonging to closed part-of-speech categories (prepositions, conjunctions, pronouns, particles) and a pre-selected list of morphological particularities, such as foreign proper names, homonymous verbs with identical lexical features and different inflections (e.g. *stati* 'to stand/to cost'), masculine nouns that inflect for (in)animacy in accusative singular (e.g. *delfin* 'a dolphin/the butterfly stroke'), lemmas with irregular or variant inflections (e.g. *a child*), and so on. The remaining and majority of the lemmas were then selected from

the list of most frequent lemmas in the then reference corpus of written Slovenian FidaPLUS, containing 620 million words (Arhar and Gorjanc 2007).

In the second stage of Sloleks compilation, lemmas were assigned their inflected forms using a program for semi-automatic paradigm generation, developed by Amebis d. o. o. for the construction of the ASES lexical database (Arhar and Holozan 2009) and related languages tools. The Sloleks morphological lexicon thus includes almost 2,800,000 inflected forms, with a quantitative description per part-of-speech category given in Table 1.

Table 1: Number of lemmas and inflected forms in the Sloleks morphological lexicon v1.2.

Part-of-speech	Number of lemmas	Number of inflected forms
nouns	54,260	924,268
adjectives	26,612	1,571,970
verbs	10,242	260,826
adverbs	6,906	9,931
numerals	2,240	18,448
pronouns	169	6,182
prepositions	96	101
interjections	85	85
abbreviations	70	70
particles	68	68
conjunctions	54	54
multi-word units[1]	3	3
TOTAL	**100,805**	**2,792,006**

2.1.2 JOS Annotation Scheme

Grammatical information in the Sloleks morphological lexicon is based on the morphosyntactic specifications developed within the "Linguistic Annotation of Slovene" (JOS) project (Erjavec and Krek 2008)[2] aimed at annotating corpora to be used in human language technologies for Slovenian. The JOS annotation scheme is based on previous projects dealing with formal grammars of Slovenian, in particular the MULTEXT (Ide in Véronis 1994) and MULTEXT-East projects (which includes most Slavic languages), with the Slovenian MULTEXT-East 4.0 specifications being identical to the JOS specifications.

1 Multi-word entries in the current version of the lexicon have been included as part of its demo integration into the *Slogovni priročnik* online style guide (Krek et al. 2013a).

2 http://nl.ijs.si/jos/index-en.html

JOS specifications include 12 part-of-speech categories: noun, adjective, verb, adverb, pronoun, numeral, preposition, conjunction, particle, interjection, abbreviation and residual, with the latter not being used in the lexicon. With the exception of particles, interjections and abbreviations, most part-of-speech categories incorporate additional grammatical features, however, not all items belonging to a particular part-of-speech category necessarily display all possible features. The list of all possible combinations of part-of-speech categories, morphological features (attributes) and their values is given in the form of a precompiled tagset[3] containing 1,902 morphosyntactic tags, while specific guidelines for their assignment to words in context are described in the corresponding annotation guidelines (Holozan et al. 2008).

As Erjavec et al. explain in more detail in their chapter in this volume, the JOS morphosyntactic specifications have primarily been developed to facilitate the development of human language technologies for Slovenian, and thus sometimes differ from the traditional grammatical descriptions given the limitations of automated natural language processing applications (Ledinek 2014a: 34–48). It is thus usually the form of a word that influences its part-of-speech classification, rather than its syntactic function. A typical example of this principle are participles ending in *-n*, *-t*, or *-č*, which are always annotated as participle adjectives, regardless of their attributive (*ukradena denarnica* 'a stolen wallet') or predicative (*denarnica je bila ukradena* 'the wallet has been stolen') syntactic role. Similar simplifications have also been implemented with specific morphological features, where, for example, the *person* feature is assigned to present tense verbs (even if they are impersonal, e.g. *dežuje* 'it rains'), and the *definiteness* feature is assigned to all adjectives (even if possessive adjectives do not inflect for definiteness).

Implicitly, through the process of manual corpus annotation and compilation of the morphological lexicon, the JOS annotation guidelines also specify the basic principles for determining the base form (lemma) of inflected word forms. These principles mostly conform to the general lemmatization principles used in other existing Slovenian language resources, e.g. selecting the nominative singular for nouns, infinitive for verbs, positive indefinite masculine singular for adjectives or word numerals, and positive for adverbs, with a few irregularities.[4] The only exception are pronouns, for which the lemma depends on the type of pronoun and its lexical features (e.g. lemmas *vame, zame, čezme* etc. for accusative bound personal pronouns inflected for number, person and gender; or the lemma *se* for reflexive personal pronominal forms *sebe/se, sebi/se, sabo/seboj*).

3 http://nl.ijs.si/jos/msd/html-sl/msd.index.msds.html

4 For example, lemmatization with nominative plural for *pluralia tantum* nouns (*alimenti* 'alimony') or the only possible form (e.g. the noun *poštev* 'account' that is only used in accusative singular as part of the multi-word expression *priti v poštev* 'to take into account'). With adverbs, the comparative (*bolj, manj, prej, raje, več, večkrat*) and superlative (*najbolj, najmanj, najprej, najraje, največ, največkrat*) forms of some adverbs represent separate lexicon forms with separate lemmas due to their specific syntactic roles.

2.2 Format

To ensure wide usability of a costly language resource, such as the reference collection of inflectional, derivational and other morphological information about the Slovenian language, it is essential to publish it as an open-source resource and encode it in a standardized way that enables flexible data organisation, as well as data comparability across databases and languages.[5] The Sloleks morphological lexicon is thus encoded as an XML document using the Lexical Markup Framework (LMF) scheme, an international standard for encoding natural language processing lexicons and machine readable dictionaries (ISO 24613:2008), developed as a common model for the creation and use of mono- and multi-lingual lexical resources, to manage the exchange of data between and amongst these resources, and to enable the merging of large number of individual electronic resources to form extensive global electronic resources (Francopoulo et al. 2006: 1).

The LMF format consists of two main types of components, the *core package* and the *extensions* of the core package. The core package defines a structural skeleton, which describes the basic hierarchy of information in a lexical database, such as information on the language, the name and accessibility of the resource (the metadata of the lexicon), as well as information on the basic structure of a lexical entry, whereas extensions give further specifications on how to combine the core package components with additional components required for a specific lexical resource, such as a morphological lexicon.[6]

The adjustment of the LMF format for standardised encoding of morphological lexica for morphologically rich languages, which has been used as the basis for encoding Sloleks, is explained in Krek and Erjavec (2009), while the full list of possible XML elements, attributes and values, together with the description of their hierarchical structure, is given in the corresponding Document Type Definition (DTD) intended for the validation of the lexicon structure.

2.3 Lexicon Entry

The basic building block of Sloleks is the lexicon entry.[7] One lexical entry consists of the lemma and its inflectional paradigm, i.e. the full list of one or more

5 The first open-source morphological lexicons were encoded in a tabular format, which is inconvenient for storing information on variant inflected forms or pronunciations, and their complex relationships with other types of information.

6 While extensions define the expected types of information in a particular lexical resource type, their number and hierarchal organisation, they do not define their semantic content, as the standardised sets of categories used for linguistic descriptions, such as the standardised names of part-of-speech categories, features and values, are defined by the ISOcat Data Category Registry (http://www.isocat.org/).

7 Although the term 'lexical entry' is used more frequently, we use the term lexicon entry to differentiate entries in a morphological lexicon from those in other types of lexical databases with prevailing semantic information.

inflected forms with corresponding grammatical information. By default, each lexicon entry includes information on lemma, its part-of-speech category and at least one inflected word form,[8] while an optional array of additional inflected forms and other morphological information is added depending on the part-of-speech and lexical features of the lemma. In the following section, we briefly present the types of morphological information found in Sloleks, their hierarchal organisation and their XML LMF exemplification.

2.3.1 Entry Key

The lexicon entry key is defined as a unique identifier used for distinguishing individual lexicon entries, since a particular lemma (the headword) can appear in several lexicon entries, either with different part-of-speech categories (e.g. the adverb and the particle *ravno* 'straight/just', the adverb and the noun *stran* 'away/page', the adverb and adjective *spet* 'again/tied') or within the same part-of-speech category (e.g. the perfective and imperfective verbs *zlagati* 'to lie/to fold', the participial and common adjective *poročen* 'married/marital', the feminine and masculine noun *prst* 'soil/finger'). Even though the entry key is primarily intended for machine processing purposes and not end-user visualisation, it is nevertheless designed so as to encode information on the part-of-speech category abbreviation and the lemma (a talking code), e.g. *S_avtomobil* for the noun 'car'. Whenever there are several identical lemmas within a part-of-speech category, an additional number identifier is added, e.g. *G_vesti_1* for the verb 'to embroid' and *G_vesti_2* for its homonymous verb 'to behave'.[9]

```
<LexicalEntry id="LE_ebc318126ea71205d05cd0ce85f86362">
          <feat att="ključ" val="R_pazljivo"/>
```

Figure 1: The entry key of the adverb *pazljivo* 'carefully' in the XML LMF format.

2.3.2 Lemma

The pivotal element of a lexicon entry to which all other types of morphological information within an entry attach is the lemma, or the entry headword. In the

8 In this paper, the terms inflectional paradigm and inflected word form are also used to describe one-word paradigms of non-inflecting part-of-speech categories, such as prepositions, as they are formally encoded in the same way.

9 Masculine and feminine pairs of surnames form a special category, as their entry key consists of information on gender instead of a number, e.g. *S_Novak_m* for male surname and *S_Novak_ž* for female surname. When a surname is homonymous with another noun of the same gender, the respective entry keys are extended by an additional number identifier, e.g. *S_Pavlica_ž_1* (for the indeclinable female surname *Pavlica*, and *S_Pavlica_ž_2* for the declinable female name *Pavlica*).

Sloleks morphological lexicon, the lemma is defined as the abstract canonical or citation form of a lexical item that unites all inflected forms with the same lexical and formal properties, and usually also the same meaning. The principles for determining entry headword in Sloleks follow the JOS lemmatization principles used in manual lemmatization of the training corpus ssj500k (Holozan et al. 2008) and the development of a data-driven morphosyntactic tagger and lemmatizer for Slovenian (Grčar et al. 2012).

```
<Lemma>
     <feat att="zapis oblike" val="pazljivo"/>
</Lemma>
```

Figure 2: The lemma of the adverb *pazljivo* 'carefully' in the XML LMF format.

2.3.3 Part-of-speech and lexical features

In addition to the obligatory grammatical information on the part-of-speech category, most lexicon entries include one or more additional lexical features, i.e. grammatical features that are assigned at the lemma-level and belong to all word forms in its inflectional paradigms, such as type (common, proper) and gender (masculine, feminine, neutral) with nouns, type (main, auxiliary) and aspect (perfective, progressive, biaspectual) with verbs, case with prepositions, and so on. Like all other grammatical features in the lexicon, lexical features are given in the form of pairs of attributes (e.g. *gender* with nouns) and their values (e.g. *masculine, feminine* or *neutral*).

```
<feat att="besedna_vrsta" val="prislov"/>
<feat att="vrsta" val="splošni"/>
```

Figure 3: Lexical properties (type = general) of the adverb *pazljivo* 'carefully' in the XML LMF format.

2.3.4 Inflectional paradigm

General information on the lexicon entry is followed by the inflectional paradigm, consisting of one or more inflected forms with corresponding information on specific grammatical features, usage frequency and compliance with the language standard (in case of variant inflected forms).

2.3.4.1 Inflected forms

In the case of uninflected par-of-speech categories, the inflectional paradigm[10] of a lexicon entry usually includes only one form, whereas the number of inflected word forms for other categories depends on the category itself, its lexical features and the degree of variability in language usage. Among the inflected part-of-speech categories, the shortest paradigms appear with adverbs and some pronouns, while adjectives display the largest paradigms, as they inflect for gender, degree of comparison, number, case and definiteness, with an average of 59 different word forms per lemma (see Table 1).

2.3.4.2 Inflectional features

Each inflected form is assigned a set of inflectional grammatical features. In contrast to lexical features, inflectional features distinguish individual forms in the inflectional paradigm of a lemma, and are therefore assigned at the level of (abstract) grammatical word forms, such as gender, number and animacy with nouns; degree of comparison with adverbs; form, person, number, gender or negation with verbs, etc. The set of inflectional features in Sloleks is based on JOS morphosyntactic specifications. However, it is not obligatory for all possible inflectional features within a part-of-speech category to be assigned to all lemmas belonging to the category, as their actual selection depends on the lemma and its lexical features.

At the same level, the lexicon also includes a mapping of all grammatical features to a position-based compact string encoding, the so-called morphosyntactic description (MSD) used in automatic morphosyntactic tagging of text corpora (see the chapter by Erjavec et al. in this volume).[11]

```
<WordForm>
     <feat att="stopnja" val="primernik"/>
     <feat att="msd" val="Rsr"/>
     /.../
</WordForm>
```

Figure 4: Inflectional features and the MSD of a comparative form the adverb *pazljivo* 'carefully' in the XML LMF format.

10 The expression "inflectional paradigm" is used to denote all the inflected forms of the lemma, as determined by the JOS system, regardless of whether they are the result of morphological (e.g. declension) or formational (e.g. gradation) processes.

11 All the comparative forms of adverbs are therefore given the "Rsr" MSD, since – in accordance with the morphosyntactic specifications of JOS – the first letter of the MSD contains the part-of-speech (R: adverb); when dealing with adverbs, the second letter then indicates the type (s: general), and the third one the degree (r: comparative).

2.3.4.3 Variants

When a given set of grammatical features (an abstract grammatical form) is realized with more than one spelling, we consider these competing word forms to be inflectional variants. They are further distinguished by the so-called variant features, which currently include information on compliance with the current language norm (as set out in *Slovenian Orthography*, 2001). Inflected forms without any normative information are considered to be in compliance with the norm (e.g. the inflected form *gradu* of the lemma *grad* 'castle' in dative singular), while the "nestandardno" attribute value denotes incompliance with the norm (e.g. the inflected form *gradi* in nominative plural). If there is a variation between two or more standard forms, they are each assigned the "variantno" label (e.g. the forms *grada* and *gradu* in genitive singular).

2.3.4.4 Corpus frequency

In Sloleks each inflected word form is also assigned its frequency in the reference 1.2 billion-word Gigafida corpus, which has been extracted automatically by querying the frequency of occurrence of the combination of the given inflected form, its lemma and its MSD. The overall accuracy of the reference morphosyntactic tagger and lemmatizer used in the annotation of Gigafida (Grčar et al. 2012) is currently 91.34 %, but varies significantly depending individual types of lemmas or word forms (ibid: 92–94).

```
<FormRepresentation>
     <feat att="zapis_oblike" val="pazljiveje"/>
     <feat att="norma" val="variantno"/>
     <feat att="pogostnost" val="97"/>
</FormRepresentation>
<FormRepresentation>
     <feat att="zapis_oblike" val="pazljivejše"/>
     <feat att="norma" val="variantno"/>
     <feat att="pogostnost" val="2"/>
</FormRepresentation>
```

Figure 5: Variant comparative inflected forms of the adverb *pazljivo* 'carefully' with normative and corpus frequency information in the XML LMF format.

2.3.5 Related forms

In addition to information on the inflectional properties of a lemma, Sloleks also includes information on its derivational connection with other lemmas or lexicon entries. The current list of derivational relations in Sloleks includes the following reciprocal relations: between a noun and its derived possessive adjective (*kruh* 'bread' and *kruhov* 'of bread'), between a verb and its gerund (*briti* 'to shave' and *britje* 'shaving'), between an adjective and a derived noun ending in *-ost* (*zarjavel* 'rusty' and *zarjavelost* 'rustiness'), between a verb and its adverbial participle (*začeti* 'to start' and *začenši* 'starting'), between a verb and its adjectival participle (*ujeti* 'to catch' and *ujet* 'caught'), between an adjective and the derived adverb (*navihan* 'mischievous' and *navihano* 'mischievously'), between an adjective and its elative (*lep* 'beautiful' and *prelep* 'too_beautiful, ~magnificent'), between an adverb and its elative (*glasno* 'loudly' and *preglasno* 'too_loudly') and between a lemma and its abbreviation (*gospod* 'mister' and *g.* 'Mr.').

```
<RelatedForm>
     <feat att="idref" val="LE_64ba3adcc4c42841599358c8
     6b738f1c"/>
     <feat att="besedna_vrsta" val="pridevnik"/>
     <feat att="lema" val="pazljiv"/>
</RelatedForm>
```

Figure 6: Related form (adjective) of the adverb *pazljivo* 'carefully' in the XML LMF format.

To summarize the above description of the Sloleks lexicon entry structure, Figure 7 shows the full set of information included in the lexical entry of the adverb *pazljivo* 'carefully', schematized to better visualise the hierarchical organisation of the original data in the XML LMF format.

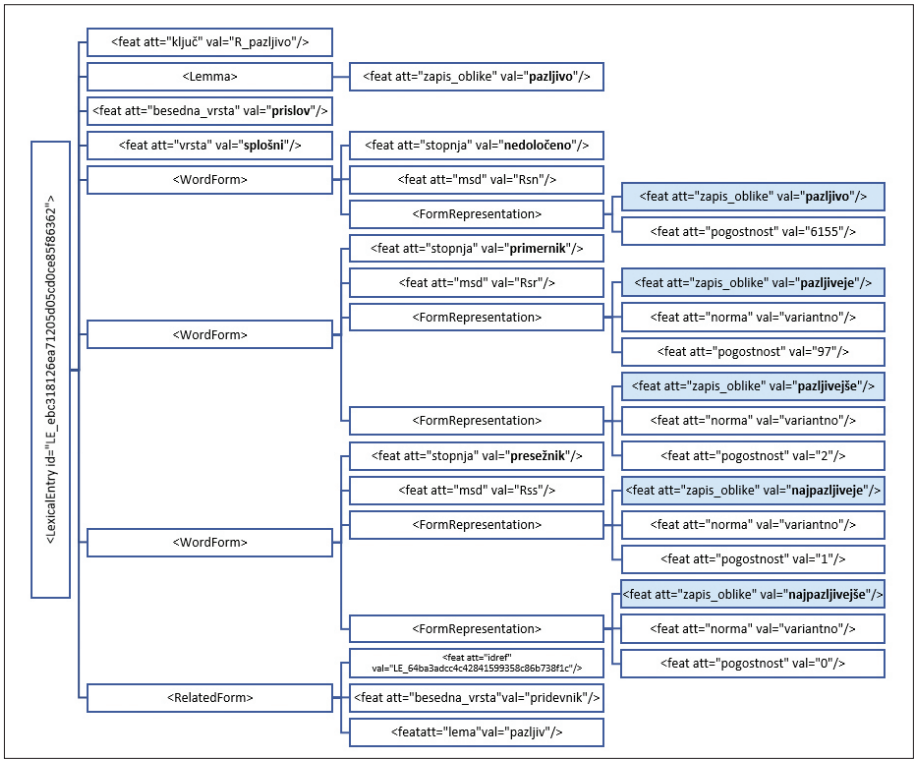

Figure 7: A schematic illustration of the full lexicon entry for the adverb *pazljivo* 'carefully' in the XML LMF format with inflected forms shaded in blue.

2.4 Visualisation

In addition to being used in various natural language processing applications, a structured collection of morphological information on Slovenian lexica that enables flexible modifications of the information that is displayed, and how it is visualised, represents an equally valuable language resource to be used as an autonomous inflection manual or integrated into other language resources, such as an online dictionary (see Dobrovoljc in this volume). An example of Sloleks lexicon visualisation has also been proposed as part of the Communication in Slovene project portal.[12]

As can be seen in the example of the visualisation of the lexicon entry for the adverb *pazljivo* 'carefully' (Figure 7) in Figure 8, the red-coloured lemma is followed by information on the part-of-speech category, lexical features and the overall

12 http://www.slovenscina.eu/sloleks

corpus frequency (a sum of frequencies for individual inflected forms included in the original database). This is followed by a separate display of the inflectional paradigm with corresponding grammatical and normative features, where specific combinations of inflectional features (grammatical forms) are separated by a line. Numbers in the frequency column include a hyperlink to the usage examples in the online corpus concordancer (corpus queries are generated automatically for the given word form, lemma and MSD combination). The bottom of the entry includes information on potential related lemmas (and their part-of-speech category), also in the form of a hyperlink to the corresponding lexicon entry.

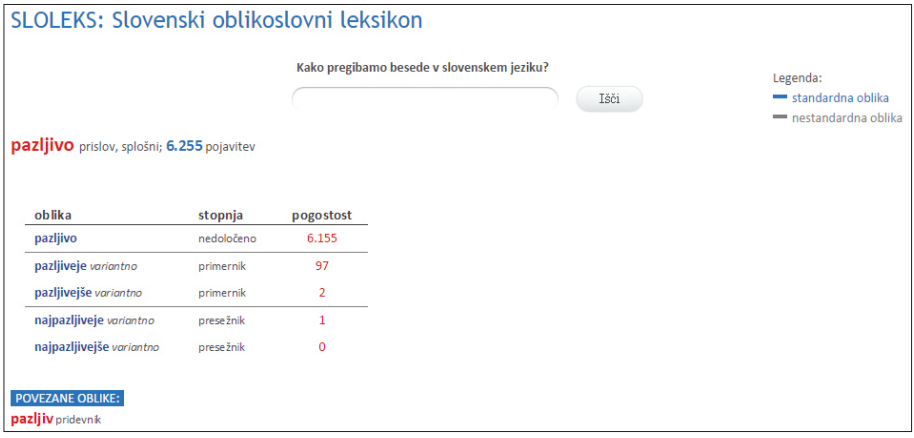

Figure 8: Visualisation of the lexicon unit for the adverb *pazljivo* 'carefully' in the Sloleks web service.

3 GUIDELINES FOR FUTURE DEVELOPMENT

3.1 Expanding the entry list

As already stated in section 2.1., the Sloleks Morphological Lexicon currently contains around 100,000 of the most commonly used lemmas in Slovene vocabulary. Compared to the glossaries of other accessible morphological resources for Slovene, which are either smaller in size (Apertium, MULTEXT-East) or are not corpus-based (SP 2001, DSLL), it currently covers the largest percentage of general Slovene vocabulary. However, the planning of dictionary and other linguistic descriptions of modern Slovene on the one side, and the growing and diverse needs for its machine processing on the other, also necessitate its further expansion. This process is envisaged as three concentric circles, each representing the fundamental starting point of the next, although not necessarily drawing on the same methodological considerations.

Given that priority is given to the integration of morphological data into a digitally-born descriptive dictionary of modern Slovene, the first concentric circle of the further expansion of the Sloleks Morphological Lexicon represents its harmonization with the dictionary database entry list, i.e. the inclusion of the (missing) core lexical units of the Slovene language, including multiword dictionary headwords, spelling variants, and other lemmas or forms morphologically linked to the lemma of a given dictionary headword.

The second circle of expansion includes the vocabulary taken from the reference corpus of the Slovene language. Although some of the reference corpus vocabulary will not necessarily become part of the dictionary, depending on the dictionary headword selection criteria, it nevertheless forms an indispensable part of various language technologies – including those used in dictionary compilation – since the lemmatizers, morphosyntactic taggers, and lexical data extraction tools must be capable of correctly recognizing both headwords and their surrounding vocabulary. In accordance with the virtuous circle of linguistic annotation, the expansion of the lexicon improves the language model of the tools, which in turn improves the accuracy of corpus annotation.

By comparing the overlap of word forms (token types)[13] in the Sloleks Morphological Lexicon with the vocabulary in the Gigafida reference corpus, we find that Sloleks contains only 43% of all token types with a minimum frequency of five occurrences in the Gigafida corpus. As expected, this share increases by increasing the frequency threshold; however, the Morphological Lexicon still covers only 79% of the total 251,292 token types that appear at least 100 times in the corpus. Such frequency of an individual token type (i.e. word form, not lemma) in a balanced and representative corpus is already a strong indicator it should be formally described in an adequate morphological database.

A more detailed analysis of the list of the most common word form types in the Gigafida corpus not yet present in the Sloleks Morphological Lexicon indicates that the database would benefit from being expanded with the following vocabulary groups:

- various types of abbreviations (*p., s., j..; nan., dok., mr.; m2, cm3, a3; UV, MMS, VIP, SUV; VPS, SŽ,* etc.);
- borrowed nouns (*city, miss, fax, art, dj, bluetooth, mac, facebook, prix, alias, maestro, college, gay, styling, fitness, volley, weekend, hiphop*, etc.);
- non-inflected attributes (*turbo, online, anti, stereo, retro, audio, etno, latino, afro*, etc.);

13 In doing so, we intentionally compared only word forms written in lowercase letters, since we did not want to depend on the automatically added data about the lemma or the spelling particularities found in corpus texts (e.g. *slovenija, ljubljana*, etc.).

- non-standard word form spellings (*tud, kr, blo, brezveze, dobr, nevem, kao, jst, jap, tolk, nč, lahk, drgač, al, tm, zarad, mislm, pomoje, una, brezveze*, etc.);
- interjections (*živjo, bognedaj, jao, jp, hehe, he, hahaha, hahahaha, sviš, hehehe, khm*, etc.);
- foreign and Slovene proper nouns (*obama, ilirika, evroliga, barca, clio, patria, beverly, pomurec, messi, airways, michel, svena, sarkozy, coca, evrovizija, titanik, čedad, Wikipedia*, etc.);
- dialect or field specific vocabulary (*škrinja, zalüblenih, mojoga, škürec, zadvečerek, špas*, etc.);
- some commonly used vocabulary or loanwords (*drugouvrščen, mimoidoči, prida, kapitalov, superpokal, štoparski, fotogalerija, tričetrt, bogve, drugoligaški, didžej, avtohiša, enoprostorec, osemvaljnik, supermodel, drska, preska, četrtinski, požarnik, klaviaturist, klientelizem, kapetanski, avtoprevozništvo, označba, predizbor, napak, prismučati, nezemljan, brezplačnik, evroobmočje, streljaj, dvetretjinski*, etc.).
- For the purpose of natural language processing, frequently used foreign vocabulary should also be recorded, such as lexical items constituting foreign proper nouns (e.g. *the, of, and*, etc.).

After expanding the lexicon with the missing headwords from the dictionary and the frequent vocabulary found in the reference corpus, the third circle of expansion foresees the inclusion of specialized vocabulary for the requirements of specific language manuals or technological applications, such as typically spoken vocabulary, vocabulary from individual areas of expertise, dialect vocabulary, or other types of vocabulary from different registers. As opposed to the first two circles, which represent the universal core of a language's lexicon description, the third circle of expansion of the lexicon cannot be foreseen or guaranteed in advance; however, it is of key importance that the community be allowed to carry out the expansion independently, by providing it with the tools and sources necessary for such task – starting with an open source database of inflectional patterns for Slovene, as discussed in the following section.

3.2 Revising morphological patterns

One of the most important tasks linked to both the expansion and re-evaluation of existing lexicons for Slovene is the creation of a finite set of machine-readable

inflectional patterns for the language, which would enable the validation of inflectional paradigms of headwords in existing reference books, the assignment of paradigms to new lemmas, and the development of methods for their automatic recognition in text corpora (e.g. Šnajder 2013 for Croatian). When we look at the range of morphological lexicons for Slovene, one could assume that there already exist several similar collections of inflectional patterns. However, these are not available to the research community at large, and the principles behind their design, classification and compliance with actual language use are mostly not documented. What is more, the initial attempts to implicitly register the complete list of patterns based on the comparison of patterns available in larger accessible reference works, such as SP2001, Apertium, and Sloleks (Dobrovoljc 2014), also revealed non-systematic pattern selection and classification, as many errors, inconsistencies or incompatibilities with contemporary language usage were identified in all three language resources.

This confirms that any upgrade or further application of the existing morphological databases in Slovenian should also involve the creation of an updated, freely accessible list of formalized inflectional patterns for the language. However, in contrast to the traditional linguistic approaches to description of morphological patterns in Slovene, their use in language technologies requires the consideration of a few additional design principles. In addition to the strict separation of inflectional patterns on the one hand, and pronunciation patterns on the other (as opposed to simultaneous description of both orthographical and pronunciation changes during inflection in DSLL, see sections XXXVIII–XLIX), as well as machine-readable formalization of patterns in the form of algorithmic rules for paradigm generation – both aspects are discussed in detail by Dobrovoljc et al. (2015), and have already been implemented in the initial Sloleks design – future revisions of the existing inflectional patterns in the lexicon should mainly focus on their compliance with actual language use.

Updating morphological information based on tendencies observed in balanced and representative corpora of modern Slovene would not only ensure an exhaustive coverage of the frequently used vocabulary (regardless of its compliance with the existing codification norm), but also enable an important re-evaluation of morphological descriptions in existing reference grammar books and dictionaries, which were not based on such vast collections of authentic language use. As demonstrated by Dobrovoljc et al. (2015), who compared the DSLL2's schemes for dynamic stress and morphology with data occurring in the Gigafida reference corpus, contemporary language use reveals the inexistence of some supposedly systemic inflectional forms (e.g. the accusative dual *cerkvé* of the noun *cérkev* 'church'), as well as the unjustifiability of some theoretic presuppositions, such as the claim that the *e* comes between two sonorant consonants in the dual and plural genitive case only when the

second sonorant is *r* (*kamra: kamer*), since usage shows that *e* may be inserted even between other combinations of sonorants (e.g. *himna: himen*; *kolumna: kolumen*; *avla: avel*).

Such re-evaluations based on analysis of authentic language use are even more important from the point of view of complete paradigm attribution, i.e. the coupling of concrete lemmas with concrete inflectional patterns, where initial analysis of attributed patterns of comparison for adverbs in the Sloleks Morphological Lexicon and the SP 2001 Slovenian Normative Guide Dictionary (Dobrovoljc 2014) revealed that both reference works diverge from common language use. For example, some adverbs that demonstrate comparison by inflection in the Gigafida corpora (e.g. *smiselno, preudarno, poredko, enakovredno, korektno, športno*) are referenced without any inflectional paradigm in one or both manuals, whereas sometimes the paradigm for comparison is attributed to adverbs that do not exhibit such behaviour in common use (e.g. *arogantno, bistroumno, strahovito, zagonetno*, etc.). Even more surprisingly, such discrepancies occur in morphological patterns for exceptions, where the Slovenian Normative Guide, for example, gives the comparative forms *dražje, ožje* and *težje* for the adverbs *drago, ozko and težko* (even though comparative forms *draže, ože* and *teže* also appear in the corpus); the forms *krajše* and *kračje* are given for the adverb *kratko* (even though the second form is not present in language use); the adverb *gladko* has the forms *gladkeje, gladkejše, glaje* and *glajše* (even though *glaje* does not occur in the reference corpus and g*lajše* has only one entry), and so on.

3.3 Categorizing variation

Morphological variation, i.e. the existence of several formal possibilities of expressing the same grammatical form, is quite common in Slovene, and occurs at various linguistic levels: the spelling (*v naprej* or *vnaprej* 'ahead'), pronunciation (*/drsáuka/* or */drsálka/* 'skater') or accentuation (*upokójenec* or *upokojênec* 'pensioner') of lemmas, as well as when selecting the morphological paradigm (*Luka: Luka or Luke* or *Lukata* for inflection of the male name Luka), the spelling or pronunciation of inflected forms (*college*: *collegea* or *college* for inflection of the loanword college), or word formation (*vanilija: vanilijev, vanilijin* or *vanilin* for forming an adjective from the noun vanilla).

Given that morphological variants in the existing version of the Sloleks lexicon are listed as word forms with identical lexical or grammatical properties, we are unable to systematically distinguish between them without additional specification of the expected differences. Consequently, since morphological lexicons are used for various purposes, it would be useful to assign the differentiation (variant)

features to the individual variants, along with their systematic classification. In doing so, we must stress that this kind of classification must not be confused with normative qualification (illustrated in Section 2.4.4.3): the first denotes a user's choice within the language system, while the second entails subsequent linguistic interpretation, which largely depends on social conventions and is thus also subject to change. Both sets of information are essential to a morphological lexicon; however, since classification enables a directed recall of individual variant forms or complete variant paradigms of one or more lexical units, while the information on their normative (non-)stigmatization is a key component for integrating the lexicon into language reference books, and can also be of value to language technology applications for text generation, such as machine translation software or speech synthesizers, that can benefit from information on the (non-)standard nature of individual variant choices.

The first attempts at systematic classification and normative qualification of variant morphological forms have already been made when establishing the design and workflow of the "Slogovni priročnik"[14] (Online Style Guide) web portal. The portal is intended as an online service for solving the most common language-related issues in Slovene text production, by juxtaposing information about the valid orthographic standard on the one hand and corpus data on the other (Krek 2012c; Krek et al. 2013a; Dobrovoljc and Krek 2013). The back-end mechanism, which connects the user's question to the relevant issue and its explanation (by visualizing the corpus and normative information for the exact queried word form(s)), takes all the necessary data from the Sloleks Morphological Lexicon, where lemmas or inflected forms, related to the language issue in question, have been adequately categorized. Each form (base or inflected) is therefore ascribed three types of categorization data: (i) the category of the issue, i.e. the type of morphological variation, which is based on an ontologically organized list of language-related issues in Slovene (Dobrovoljc and Krek 2011; Bizjak Končar et al. 2011), (ii) the type of variant within the category, and (iii) its normative value.

An example of such a categorization is shown by a fragment of the lexicon unit for the noun *Klemen* in Figure 9. At the first level, the lexical unit is already carrying the information about its link to language issue no. C1a3a (*Morphology* > *Nouns* > *Masculine Declensions* > *Nouns with Unstable Vowels* > *Slovene Proper Nouns*), while individual forms in the subsequent paradigm also include information about the specific variant they belong to (C1a3a_s_1, for example, is used to denote a paradigm that omits *e,* while C1a3a_s_2 a is used to denote a paradigm without this omission), as well as the information on their normative qualification (e.g. the *variantno* qualifier that marks a standard double).

14 http://slogovni.slovenscina.eu/

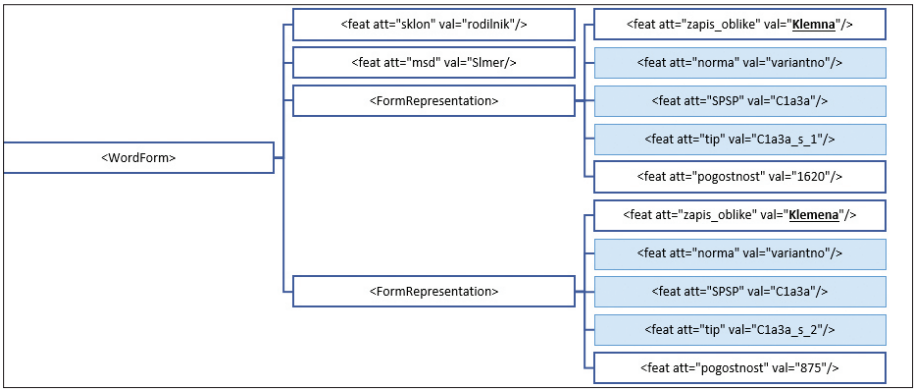

Figure 9: Part of the *S_Klemen* lexicon unit with a category of variant declension.

This kind of categorization of morphological characteristics in the database thus enables a controlled recall of data within an individual lexicon unit, such as the list of all relevant linguistic issues related to the unit, or one or more forms of a given variant paradigm. On the other hand, it also enables an automatic recall of the list of all other lemmas which display the same kind of morphological, derivational or pronunciation variance, e.g. all Slovene proper nouns with an unstable vowel.

3.4 Adding pronunciation

The Sloleks Morphological Lexicon currently does not include data on the pronunciation characteristics of the word forms it contains, meaning that the word forms included are not accented. One of the priority upgrades to the existing version of the Sloleks Morphological Lexicon is thus the incorporation of pronunciation data, with the aim of providing a comprehensive description of both inflectional and phonetic characteristics of contemporary Slovene vocabulary. This is especially important from the point of view of speech technology, since Slovenian linguistic infrastructure currently lacks a freely accessible lexicon needed for the development of speech recognizers and synthesizers for various applications, such as subtitle generators, screen readers for visually impaired, natural language interaction systems, and the like.

The pronunciation information, based on a standard machine-readable phonetic alphabet, should be included at the level of both lemmas and inflected forms. In cases of pronunciation variation, a common phenomenon in Slovene, one orthographical word form can thus have several pronunciations assigned; similarly to

dealing with the variation of non-accented, orthographical forms, we can distinguish between them by using adequate qualifiers, which allow us to automatically recall the pronunciation of an individual form or all forms in one of the variant pronunciation paradigms (see Section 3.3 Categorizing variation). This approach is used for all types of pronunciation variance, regardless of whether we are dealing with phonemic (prevajalka: *prevajalka-prevajaụ̑ka*) or accentual variance (agencija: *agéncija-agencíja*) of all or just one of the inflectional forms in a given paradigm.

Just like in the current version of the Sloleks lexicon, adding pronunciation information would not change the fact that lemmas with the same spelling and pronunciation are separated into several independent lexical units if they display different expressive characteristics, i.e. if they fall under different parts-of-speech (e.g. the adverb and adjective *spet*), have different lexical properties (e.g. the feminine and masculine noun *prst*), or different inflections (e.g. the verbs *vesti*: *vedem* and *vesti: vezem*). Similarly, no changes would apply to homonymic pairs of lexemes with identical formal, but different semantic properties (e.g. the masculine nouns *bor* 'pine tree' or *bor* 'chemical element'), which would continue to be processed as one distinct unit of vocabulary with only one corresponding lexicon unit (the masculine noun *bor*), regardless of their meaning.[15] Since the Morphological Lexicon does not record tonemic accent, the same rule applies to pairs of semantically differing homographs that are differentiated only by their tonemic accent (e.g. the adjectives *bȗčen* 'of a pumpkin' and *búčen* 'loud' thus share a common inflectional paradigm of the general adjective *bučen*).[16]

On the other hand, adding pronunciation information would change the treatment of lemmas with the same spelling, but a different pronunciation, e.g. *partija* (pronounced *partíja* 'the (Communist) party' and *pártija* 'the match') or *častiti* (*částiti* 'to buy somebody a drink' and *častíti* 'to worship'), which to now were considered a single lexicon unit due to their identical lemma, grammatical patterns and non-accented inflectional paradigms. By introducing semantically differentiating pronunciation information, both lexemes become independent lexicon units (*S_partija_1* in *S_partija_2*). However, it should be noted that current morphological analysers for Slovene do not enable semantic disambiguation of morphologically overlapping homographs within a given context, which is why word forms belonging to such homographs would be given an identical lemma and morphosyntactic tag. In turn, the corpus frequency information (see Section 2.4.4.4.) for identical word forms with identical grammatical features would be identical for both lemmas.

15 For the relationship between the lexicon and dictionary headword, see the paper by Dobrovoljc in this publication.

16 For the relationship between formally motivated lexicon units and semantically motivated dictionary units, see Gantar (2015) and K. Dobrovoljc in this publication.

4 CONCLUSION

The Sloleks Morphological Lexicon, together with its morphological, derivational, normative, distributional and other types of linguistic data, represents a common intersection point between the various language resources foreseen by the proposal for a new dictionary of modern Slovene (Krek et al. 2013b), such as reference, balanced, spoken, historical, and other types of linguistically annotated corpora. On the other hand, the data from the lexicon are equally useful in reference language manuals, such as (digital) dictionaries, online style guides, grammars and others. By introducing a systematic approach to the description and formalization of Slovene morphology, the Sloleks lexicon enables a uniform and consistent treatment of morphological phenomena within the fields of both language technologies and language resources. As such, it aims to overcome one of the key deficiencies of Slovene natural language processing and Slovene language teaching – from primary and secondary schools to teaching Slovene as a foreign language.

Future development of the lexicon should mainly focus on a significant expansion of its entry list, including multi-word units, usage-based revision of the existing morphological patterns and their attribution to individual lemmas, systematic linguistic and normative categorization of frequent morphological variation, and addition of pronunciation information. All these processes must be implemented with the current state-of-the-art technologies, many of which are already available for Slovene. The future development of the Sloleks lexicon should thus be understood as an ongoing process without a final endpoint, since languages are always accruing new words, which need to be both adequately described and efficiently processed. With this in mind, the widespread usability of the lexicon can only be assured by a continued open access to this resource. This not only justifies the investment into its development, but also gives the Slovene language the opportunity to survive in the coming digitized world.