

Morphological Information in Modern Slovene Dictionaries

Kaja Dobrovoljc

Abstract

Although morphology in lexicography is generally considered to be a solved problem which mostly deals with user-oriented evaluations of its comprehensibility, online dictionaries bring new possibilities for both dictionary users and makers alike. In the context of planning a future dictionary of modern Slovene, this paper explores the language users' need for morphological information, and the different aspects of its inclusion in a born-digital online dictionary. Preliminary analysis of inflection dictionary log files confirms that there is a great need for the inclusion of inflectional information, and that users tend to search for both regular and irregular inflectional paradigms. However, this need is not sufficiently met within the recently issued edition of the reference *The Dictionary of Slovene Literary Language*, as decoding inflectional and other morphological information requires substantial cognitive effort and metalinguistic knowledge that cannot be expected from most users. Given that Slovenian is a morphologically rich language with extensive inflectional information, we take into account the idea of a separate machine-readable morphological database intended for use in language guides and various NLP applications. This database brings many advantages for dictionary users, such as the display of full inflectional, pronunciation and derivational paradigms, normative information, hyperlinking, improved searching, corpus linking, speech synthesis and voice search recognition. At the same time, it demands careful consideration of the content-related, visual and technical issues that arise when interlinking two distinct databases, in particular morphology-dependent polysemy and variant spelling synonymy.

Keywords: morphology, inflection, morphological lexicon, dictionary database

1 INTRODUCTION

In addition to information on the semantic properties of lexical items, dictionaries usually also include information on their formal properties, such as pronunciation, inflection, orthography and so on. Contrary to the reception-based semantic description, such information advises users on how to use lexical items in the process of actual production. This has also been standard practice in Slovenian lexicography, as ever since the first edition of *The Dictionary of Slovene Literary Language* (DSL) most subsequent dictionaries have considered information on pronunciation, inflection and other morphological features as an indispensable part of a dictionary description, regardless of the dictionary type, i.e. general, specialised, terminological, historical, dialectical or any other type of monolingual dictionary.

Despite the fundamental role of morphological information in lexical descriptions of a language, however, there has been relatively little research on questions related to this particular aspect of lexicographic work, in Slovenian and general lexicography alike. While research related to (paper-based) dictionaries for morphologically less complex languages mostly discusses how much morphological information should even be included in dictionaries, in addition to irregular morphological phenomena, and to what extent can regular morphological patterns be predicted by non-native dictionary users (Jackson 2002: 105–107; Honselaar 2003: 355–356; Caluwe and Taldeman 2003: 73–77), research related to morphologically rich languages mainly focuses on the micro-structural issues of the optimal presentation of inflectional information, such as ways of abbreviating inflected forms or cross-referencing paradigmatic patterns, and the level of comprehensibility with regard to dictionary users (Vikør 2009: 140; Kola 2012). On the other hand, rather than ways of encoding morphological information, Slovenian linguistics has mainly been concerned with the question of its suitability from the viewpoint of (literary) language standardisation (see Toporišič 1971a and 1971b; Rigler 1971 and 1972).

Given the many possibilities that the online dictionaries bring to dictionary users and makers alike, the present paper aims to explore the prospects of describing and presenting morphological information in a born-digital dictionary of modern Slovene. We first perform an empirical analysis of the user needs for morphological information in Slovenian dictionaries (section 2), and investigate how these are met within the recently issued reference *The Dictionary of Slovene Literary Language, second edition* (section 3).¹ Given the general consensus of storing morphological information in the form of a separate machine-readable morphological database (morphological lexicon), we discuss the possible advantages of this approach for dictionary users (section 4.1), and also emphasize the need

1 Although this method is applicable to morphology in general, the remainder of this paper mostly focuses on inflectional information.

for a clear distinction between information in a morphological database and its presentation in a dictionary (section 4.2), as well as the distinction between a lexicon entry and a dictionary entry (section 4.3).


2 USER NEEDS

As a starting point for evaluation of user needs with regard to including morphological information in online language resources, this section presents an initial analysis of query log files of the Amebis inflection dictionary,² developed as one of the modules of the Besana grammar checking application (Holozan 2012). The demo version of this module is designed as an online dictionary portal that provides information on inflected forms (both standard and non-standard), derived forms and grammatical features of words or multi-word expressions entered by the user (Figure 1). The inflection dictionary is based on the ASES lexical database (Arhar and Holozan 2009), which is continuously developed and currently contains approximately 244,000 lexical entries.

Predstavitvena verzija pregibnika Amebis Besana 4.10.2

Za preizkus kakovosti pregibanja v polje spodaj vpišite besedo ali besedno zvezo. Če nimate nameščene slovenske tipkovnice, pišite *kaša* ali *ka'sa* in ne *kasa*, *kasha* ali *kas'a*.

Oblika za pregibanje

gospa 

samostalnik
občno ime

♀ ženski spol

	ednina	dvojina	množina
imenovalnik	gospa	gospe	gospe
rodilnik	gospe	gospa	gospa
dajalnik	gospe <i>gospej</i>	gospema	gospem
tožilnik	gospo	gospe	gospe
mestnik	gospe <i>gospej</i>	gospéh	gospéh
orodnik	gospo	gospema	gospemi <i>gospami</i>

svojljni pridevnik
gospeljn

Figure 1: An example of the Amebis Besana dictionary entry for the noun *gospa* 'lady/Mrs' (Slovenian interface only).

² <http://besana.amebis.si/pregibanje/>

Our analysis is based on an extensive log file for a six-year period from January 2009 to January 2015, which has been compiled as a two-column list of distinct query strings (words or multi-word expressions entered by the user) and the number of such queries. As can be seen in Table 1, 2,350,778 queries of 787,751 distinct query strings were recorded in this period. Thus, on average, more than 1,000 queries were recorded daily,³ which confirms the significant need for this type of linguistic information by users, especially given that Besana is only one of several freely available online inflection dictionaries for Slovenian.⁴

Table 1: Number of queries within the Amebis Besana dictionary in the period 2009–2015.

Type of query string	Number of queries	Distinct query strings
Word	2,250,705	723,608
Multi-word expression	100,073	64,143
TOTAL	2,350,778	787,751

To gain a better understanding of which lexical items users investigate most frequently and in what way, we limited the subsequent qualitative analysis to query strings occurring in 300 or more queries. Even though these include only 571 distinct strings, they represent more than 25% of all queries (619,117 queries in total), which signals that speakers of Slovenian find the inflection of some lexical units significantly more problematic than others.

The results given in Table 2 show that these mostly include common nouns, such as *hiše*⁵ (English ‘houses’; 26,115 queries), *otrok* (‘child’; 22,164), *dan* (‘day’; 15,488), *hči* (‘daughter’; 14,046), *mati* (‘mother’; 10,824), *gospa* (‘lady’; 10,756), *človek* (‘man’; 6,941), *tla* (‘floor’; 6,006), *otroci* (‘children’; 4,838), *vodja* (‘leader’; 4,782), *pljuča* (‘lungs’; 4,501), *vrata* (‘door’; 4,408), *drva* (‘wood’; 4,199), *oko* (‘eye’; 4,034), *hiša* (‘house’; 3,957), *dno* (‘bottom’; 3,333), *pes* (‘dog’; 3,296), *breskev* (‘peach’; 3,032), *okno* (‘window’; 2,991), and *leto* (‘year’; 2,967). These are followed by verbs, such as *zvedeti* (‘to find out’; 4,426), *dati* (‘to give’; 3,401), *biti* (‘to be’; 3,394), *iti* (‘to go’; 3,201), *jesti* (‘to eat’; 2,259), *imeti* (‘to

3 As a point of comparison, the online portal for the reference Slovenian orthography guide (<http://bos.zrc-sazu.si/sp2001.html>) recorded an average of 400 queries daily in the period from March 2010 to June 2015. In their overview of the frequency of usage for different online dictionaries, Bergenholtz and Johnsen (2005: 122–126) report on a range from a few hundred to a few thousand queries per day, for languages or language combinations with a considerably higher number of speakers than the two million seen for Slovenian.

4 A similar type of full paradigm querying is offered by the Sloleks morphological lexicon interface (available as part of the <http://eng.slovenscina.eu/sloleks> and <http://www.termiana.net> portals), while abbreviated inflectional information is also included in most of the dictionaries produced by the Fran Ramovš Institute of the Slovenian Language (available as part of the www.fran.si, <http://bos.zrc-sazu.si/> and <http://www.termiana.net> portals).

5 The list of most frequent queries presented in this paper does not exclude queries suggested as demo queries or those used in system testing, such as *hiše* ‘houses’, *hiša* ‘a house’, or *Oselica* (name of a village).

have'; 1,736), *vedeti* ('to know'; 1,474), *moči* ('to be able'; 1,100), *delati* ('to work'; 1,090), *poslati* ('to send'; 1,076); pronouns, such as *on* ('he'; 4,325), *nič* ('nothing'; 3,518), *jaz* ('I'; 3,334), *ta* ('this'; 3,292), *ona* ('she'; 3,059), *kaj* ('what'; 2,473), *kar* ('which'; 2,205), *kateri* ('which'; 2,079), *ti* ('you'; 1,901), *moj* ('my'; 1,892); and proper nouns, such as *Oselica* (20,731), *Miha* (5,271), *Luka* (5,120), *Marko* (3,115), *Jaka* (2,310), *Žiga* (2,144), *Mitja* (2,046), *Grosuplje* (1,985), *Sašo* (1,722), *Klemen* (1,598). There is significantly less recorded queries for adjectives, e.g. *lep* ('beautiful'; 1,183), *nov* ('new'; 690), *dober* ('good'; 686), numerals, e.g. *dva* ('two-masculine'; 2,530), *tri* ('three'; 1,500), *dve* ('two-feminine'; 921), and adverbs, e.g. *lahko* ('easy'; 685), *dobro* ('well'; 593), *rad* ('gladly'; 562), which suggests users find these less problematic due to their regular inflectional patterns. The analysed list does not include any multi-word expressions, as even the most frequently queried multi-word unit (*dve leti* 'two years') does not reach the threshold, with only 241 queries in total.

Table 2: The list of most frequent queries per part-of-speech category in the Amebis Besana inflection dictionary.

	distinct queries	all queries
common nouns	336	398,658
verbs	71	53,633
pronouns	64	72,247
proper nouns	63	66,681
adjectives	14	6,878
numerals	10	9,003
adverbs	7	3,344
other	6	8,673
TOTAL	571	619,117

As expected, the most frequently queried strings include well-known words with irregular conjugation or declension patterns, which are also frequently discussed in language-related online forums (Dobrovoljc and Krek 2011; Bizjak Končar et al. 2011) and amongst the most common mistakes in student essays (Kosem et al. 2012a). On the other hand, our query log analysis reveals a surprisingly high number of queries related to seemingly unambiguous words, which inflect by regular patterns and have thus not been given any special consideration in existing language manuals so far, such as *avto* ('car'; 2,034), *mama* ('mother'; 1,578), *miza* ('table'; 1,565), *stol* ('chair'; 1,319), *fant* ('boy'; 1,070), *ura* ('clock'; 922), *knjiga* ('book'; 918); *delati* ('to work; 1,090), *videti* ('to see'; 776), *hoditi* ('to walk'; 744), *govoriti* ('to talk'; 612), *dobiti* ('to get'; 494); *lep* ('beautiful'; 1,183), *nov* ('new'; 690), *prvi* ('the first'; 447), *star* ('old'; 368), and *zanimiv* ('interesting'; 309).

Even though the original log files lack other potentially relevant metadata on individual search queries, such as user ID, user demographics or look-up duration, which could give better insights into the user profile and the relevance of the obtained results (see for example the Wiktionary log files used in Müller-Spitzer et al. 2015), the results of this elementary query log analysis nevertheless illustrate there is a significant need to include inflectional and other morphological information in future lexical descriptions of Slovenian, and at the same time indicate this need is not limited to a closed set of well-known exceptions, but also includes lexical items with regular inflection.

3 MORPHOLOGICAL INFORMATION IN DSLL2

In the introduction section, the authors of the second, revised and partially updated edition of the *Dictionary of Slovene Literary Language* (DSLL2), the reference dictionary of standard Slovenian, describe the dictionary as a source of information on both semantic and formal properties of Slovenian lexica, since “for each word, the dictionary explains how it is written and pronounced, what are its dynamic and pitch accents, how it inflects, what it means and what are the relations between individual meanings” (Gliha Komac et al. 2014: 25, translated by K. D.). In both printed and online versions, DSLL2 continues the tradition of the first edition (DSLL, issued in 1970–1991), in which the information on inflection is presented as a combination of abbreviations in the dictionary entry, with instructions on how to interpret these in the dictionary’s introduction. In order to access information on inflection of a lexeme, the dictionary users therefore first need to know how this information is encoded and then familiarize themselves with specific decoding instructions in the introduction section for their appropriate interpretation. In general, this can be described as a four-stage process consisting of (i) identification of the headword (DSLL2 Introduction: §27–§29), (ii) identification of the headword part-of-speech category (§30), (iii) decoding of the second/third basic form (§160–§165), and (iv) classification into the appropriate pattern for inflection and stress (§180–§196).

Although the initial phase of identifying the relevant headword seems relatively trivial, the results of the log file analysis presented in Section 2 show that users often query non-canonical word forms, which is why retrieving inflectional information from a dictionary should not be conditioned on comprehending the lemmatization principles used for headword selection. In addition to querying ambiguous inflected forms, such as *gospe* (inflected form of ‘lady’), *hčer* (inflected form of ‘daughter’), *dni/dnevi* (inflected forms of ‘days’), *njih* (‘them’), *matere* (inflected form of ‘mother’), *brki* (plural form of ‘moustache’), *starš*

(singular form of collective noun ‘parents’), or *sabo/seboj* (instrumental form of ‘oneself’), the list of most frequent queries in the Amebis Besana dictionary also includes words for which we can assume the users intended to enter an abstract canonical form, but chose the ‘wrong’ (non-standard) spelling, e.g. *imati* (instead of *imeti*, ‘to have’) or *pluča* (instead of *pljuča*, ‘lungs’), or the ‘wrong’ (non-prototypical) grammatical features, such as number or gender, e.g. *psi* (‘dogs’ in plural), *smuči* (‘skis’ in plural), *dve* (‘two’ in feminine), *ona* (‘she’), *vsí* (‘everybody’ in plural), *midve* (‘us’ in feminine dual), or *onidve* (‘they’ in feminine dual). With regard to DSL2, these findings raise a particular concern with respect to its online version,⁶ as looking up lexemes in a form different than the headword, such as an inflected form or a variant spelling, only gives results if the queried string appears as part of the grammatical information slot following the headword (e.g. there are no hits for querying *pluča*, the frequent non-standard spelling of the noun ‘lungs’). On the other hand, adjusting the default settings to search through full dictionary entries, and not just the headword and its grammatical information, returns all dictionary entries containing the queried string, regardless of their relevance to the user (e.g. 78 dictionary entries for querying *psi*, the plural form of the noun ‘dog’).

Similarly, the second stage of identifying the part-of-speech category and other grammatical features of the headword needed for subsequent identification of the corresponding morphological pattern can also pose a challenge to non-professional users, as these can be given in different sections of the dictionary entry: either immediately after the headword, in the form of a qualifier with an abbreviation of the part-of-speech category or one of its features (e.g. *finale -a m (ā)* or *zanimiv -a -o prid.*, where *m* denotes masculine noun and *prid.* denotes adjective), as part of the definition (*sêstrin -a -o (ē) svojilni pridevnik od sestra* ‘possessive adjective of *sestra*’ or *bitP -ā -ó in -ò opisni deležnik od biti sem (î ä õ)* ‘descriptive participle of *biti*’), in the so-called qualifying explanation (*anglo- prvi del zvez (ā) first part of phrases*’ or *sp členica ‘-particle’*), in a separate entry (*aloha gl. aloja ‘aloha see aloja’*), or this information is simply missing from the dictionary (*kamen... prim. kamn... ‘kamen... see kamn...’* and *kamn... prim. kamen... ‘kamn... see kamen...’*). After having identified the headword and the part-of-speech category of the lexeme of interest, the user should then consult the introduction section to find appropriate instructions on how to decode the abbreviated second or third (adjectives only) form listed next to the headword, e.g. *-a* in the *finale* example above. These instructions, however, demand a relatively high level of linguistic knowledge, which cannot necessarily be expected from non-professional users or non-native speakers, for example:

Nouns and adjectival words are abbreviated in the following way: a) When the first word form ends in a consonant, the second word form is formed

⁶ <http://www.sskj2.si>

by adding the given part of the second word form, consisting of a vowel or *j, n* + vowel, to the first word form /.../. The second word form is formed in the same way, when the given part of the second word form is *-ih* /.../. If a longer part of the second word form is given or the ending is preceded by a consonant (other than *j, n*) due to changes in endings, the given form indicates which part of the headword it applies to /.../. b) When the first word form ends in a vowel, the second word form is formed by adding the given part of the second word form, consisting of *j, t, n* + vowel, to the first word form /.../, or the last vowel of the first word form is omitted, if the given part of the second word form begins with a vowel /.../. In the same way, the second word form of a noun ending in *-ega*, which is otherwise inflected by adjectival declension /.../. If the given part of the second word form begins in a consonant (other than *j, t, n*), the given form indicates which part of the headword it applies to /.../. (DSLL2 Introduction: § 161, translation by K. D.)

Another potential issue in the comprehensibility of morphological information in DSLL2 is information on inflection of the so-called cross-referencing headwords, pointing to an entry with a more standard-like spelling of the headword, when the two headwords do not inflect in the same way. For example, the entry for the word *croquis* (***croquis*** gl. *kroki* ‘croquis see croquis’ points to the dictionary entry of its spelling variant ***kroki*** *-ja m (i)*, but the two headwords have different inflectional paradigms (e.g. *kroki+ja* vs. *croquis+a* in genitive singular). What is more, some dictionary entries also lack the abbreviated second word form needed for subsequent inflection pattern deduction, such as ***múlda*** *ž (û) jarek za odtok tekočine s ceste, tlakovanih površin* or ***rímokatoličánka*** *ž (î-â) pripadnica rimskokatoliške vere*.

In the last stage of the inflection deduction process, users then use the combination of the headword and its un-abbreviated second or third form(s) to select the appropriate governing scheme for inflection and stress (Figure 2) and its specific subtype, which also requires some knowledge of linguistic terminology (e.g. *base/ending stress, stress on different base syllables, short/long stress* and so on), consideration of exceptions and modifications signalled in footnotes, and understanding the meaning of special symbols, such as the symbol ~ (denoting the formation of the inflected form based on the nominative or infinitival base form or part of the base form), the symbol – (denoting either formation based on genitive or present base form or part of the base form, or the nominative masculine or feminine form for adjectives), and the symbol ‘ (denoting the place of stress).

§ 188 SAMOSTALNIK

I. NAGLAS NA ISTEM ZLOGU IMENOVALNIKA IN RODILNIKA
A. NAGLAS NA OSNOVI
a) Moški spol⁶
1. Samostalniki s končnico -a v rodilniku

ed. im.	rod.	daj.-mest.	or.	mn. im.	rod.	daj.	tož.	mest.	or.	dv. im.-tož.	daj.-or.	
-0	-a	-u	-om ⁷	-i	-ov ⁷	-om ⁷	-e	-ih	-i	-a	-oma ⁷	
(-a)												
(-e)												
(-o)												
(-um)												
(-us)												
					-ovi	-ov	-ovom	-ëve	-óvih	-óvi	-óva	-óvoma
				-a (s)	-0	-om	-a	-ih				
rák	<i>ráka</i>	<i>ráku</i>	<i>rákóm</i>	<i>ráki</i>	<i>rákóv</i>	<i>rákóm</i>	<i>ráke</i>	<i>rákíh</i>	<i>ráki</i>	<i>ráka</i>	<i>rákoma</i>	
drvár	<i>drvárja</i>											
komité	<i>komitéja</i>											
nágelj	<i>nágeljna</i>											
fanté	<i>fantéa</i>											
slúga	<i>slúga</i>											
finále	<i>finála</i>											
máksimum	<i>máksíma</i>											
					<i>denárci</i>	<i>denárcev</i>	<i>denárce</i>	<i>denárce</i>	<i>denárčih</i>	<i>denárci</i>		
					<i>grobívi</i>	<i>grobív</i>	<i>grobívom</i>	<i>grobíve</i>	<i>grobívih</i>	<i>grobívi</i>	<i>grobíva</i>	<i>grobívoma</i>
					<i>abstráku (s)</i>	<i>abstráku</i>	<i>abstrákóm</i>	<i>abstráku</i>	<i>abstrákíh</i>	<i>abstráki</i>		

Opomba: Nemi e se v tujkah ohrani, če določa izgovor predhodnega soglasnika (*bridge* [brídz-] *bridgeu* [bríđu] proti *brumaire* [brímèr-] *brumaira* [bríméru]).

⁶ Pri samostalnikih, ki poznajo podspol človeškosti oziroma živosti, je tožilnik ednine enak rodilniku, pri drugih pa imenovalniku.
⁷ Za *c, j, č, ž, š, dž* se v končnici o premenjuje z *e*.

Figure 2: An example of a governing scheme for inflection of nouns by first masculine declension (with footnotes) in DSL2 Introduction.

The selection of an appropriate pattern can also depend on other inflected forms given in the dictionary entry (in addition to the default headword form and the abbreviated second/third form), but not always, as these can also signal particularities of an individual part-of-speech category or word forms that the lexicographer deemed to be potentially ambiguous, without any influence on the pattern deduction process (DSL2 Introduction: §184–185), although the dictionary does not specify how users can distinguish between these competing interpretations. Similarly, a full list of word forms is given for headwords that cannot be placed in one of the patterns in the Introduction, as illustrated in Figure 3.

ón óna -o stil. -ó zaim., ed. m. njêga, njêmu, njêga, njêm, njím, enklitično rod., tož. ga, daj. mu, enklitični tož. za enozložnimi predlogi -nj oziroma -enj [ənj] , če se predlog končuje na soglasnik; ž. njé, njêj tudi njèj tudi nji, njó, njêj tudi njèj tudi nji, njó, enklitično rod. je, daj. ji, tož. jo, enklitični tož. za enozložnimi predlogi -njo; s. kakor m., le tož. óno stil. onó tudi njêga; mn. m. óni stil. oní, njih, njím, njih in njé, njih, njími, enklitično rod., tož. jih, daj. jim, enklitični tož. za enozložnimi predlogi -nje; ž. óne stil. oné dalje kakor m.; s. óna stil. oná dalje kakor m.; dv. m. ónadva tudi onádva stil. óna, njíju tudi njih tudi njih dvéh stil. njú, njíma tudi njíma dvéma, njíju tudi njih tudi njih dvá stil. njú, njíju tudi njih tudi njih dvéh tudi njíma tudi njíma dvéma, njíma tudi njíma dvéma, enklitično rod., tož. ju in jih, daj. jima, enklitični tož. za enozložnimi predlogi -nju; ž. ónidve stil. onédve dalje kakor m., le tož. njíju tudi njih tudi njih dvé stil. njú; s. kakor ž. (ò ó)

Figure 3: Full inflection paradigm given at the beginning of the dictionary entry for the pronoun *on* 'he' in the online version of DSL2 (small font denotes qualifiers).

Although DSLL2 is considered to be the reference manual for inflectional and other morphological information on Slovenian lexica, it seems this purpose is not achieved in an optimal way. The four-stage inflection pattern decoding process presented above presents a challenge for dictionary users, requiring them to combine specific information from the dictionary entry and general instructions from the terminologically challenging introduction section.⁷ Although such a method of encoding morphological information is understandable given the practical limitations of the paper-based first edition of DSLL, it is less justifiable in its second edition, published more than 40 years later in both print and online versions, especially given the fact that language professionals themselves pointed out the difficult decoding of inflection, stress and pitch patterns in the DSLL2 planning discussions (Perdih 2008: 18, 136, 142–143).

4 MORPHOLOGICAL LEXICON AS A COMPONENT OF AN E-DICTIONARY

The fact that a born-digital online dictionary enables a new approach to describing and presenting morphological information for Slovenian has first been recognized by the authors of the recent “Proposal for a Dictionary of Modern Slovene” (Krek et al. 2013b), who suggest storing morphological information as part of a separate database, an enhanced version of the Sloleks reference morphological lexicon of Slovenian language (presented in the chapter by Dobrovoljc et al. in this book), and visualising it in a separate section of the dictionary entry (the so-called *Inflection* tab). A similar solution has also been proposed by the authors of the “Draft Concept of the New Dictionary of Slovene Literary Language” (NDSLL; Gliha Komac et al. 2015) who speak of a lemmatization database with information on the formal properties of lexical units that would be displayed as part of the *Pronunciation and inflection* section of the online dictionary entry.

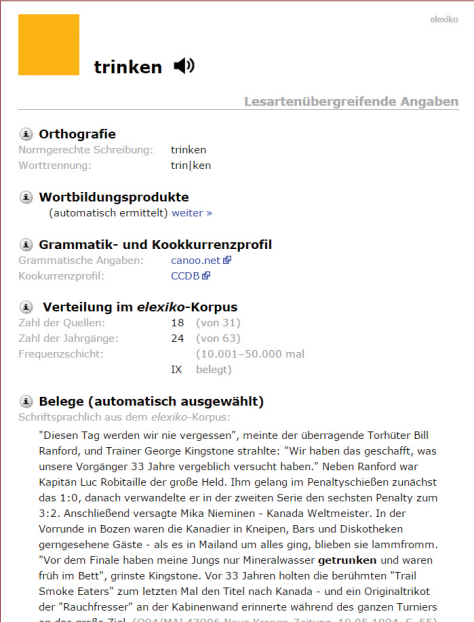
This idea of a separate, but integrated machine-readable morphological database (a morphological lexicon) has several advantages for both the quality of dictionary information and the resulting user experience (as discussed in section 4.1), but it also raises new questions on the relation between information stored in the lexicon and that shown in the dictionary (4.2), and the relation between the lexicon and the dictionary entry (section 4.2.).

⁷ This process is particularly problematic with respect to students and non-native speakers, who are believed to look up regular forms and patterns more often, since regular forms can only be decoded from the abbreviated patterns in the Introduction section, in contrast to irregular forms that are usually given in the dictionary entry itself.

4.1 Morphological lexicon as a source and navigator of dictionary information

In the context of building a future dictionary of modern Slovene, a machine-readable morphological lexicon fulfils two distinct roles. On the one hand, it is used as a key component in the development of different NLP applications for grammatical annotation of corpora and subsequent lexical data extraction (see the chapter by Erjavec et al. in this book). On the other hand, a morphological lexicon presents the primary source of information on the formal characteristics of lexical units in a dictionary, such as information on their part-of-speech category and other morpho-syntactic features, or information on their inflection, derivation and pronunciation.

In related born-digital online dictionaries for other languages, inflectional paradigms are usually presented in full and without abbreviated forms, either through a hyperlink to an external dictionary of inflected forms (as with the Icelandic ISLEX multilingual online dictionary, the BFL lexical database for French or the Elexiko dictionary portal for German illustrated in Figure 4), or as part of the dictionary entry itself. For morphologically less complex languages, the latter solution usually includes listing inflected forms, pronunciations and related grammatical information



trinken

Lesartenübergreifende Angaben

Orthografie
 Normgerechte Schreibung: trinken
 Worttrennung: trin|ken

Wortbildungsprodukte
 (automatisch ermittelt) weiter >

Grammatik- und Kookkurrenzprofil
 Grammatische Angaben: canoo.net
 Kookkurrenzprofil: CCDB

Verteilung im elexiko-Korpus
 Zahl der Quellen: 18 (von 31)
 Zahl der Jahrgänge: 24 (von 63)
 Frequenzschicht: (10.001–50.000 mal
 IX belegt)

Belege (automatisch ausgewählt)
 Schriftsprachlich aus dem elexiko-Korpus:
 "Diesen Tag werden wir nie vergessen", meinte der überragende Torhüter Bill Ranford, und Trainer George Kingstone strahlte: "Wir haben das geschafft, was unsere Vorgänger 33 Jahre vergeblich versucht haben." Neben Ranford war Kapitän Luc Robitaille der große Held. Ihm gelang im Penaltyschießen zunächst das 1:0, danach verwandelte er in der zweiten Serie den sechsten Penalty zum 3:2. Anschließend versagte Mika Nieminen - Kanada Weltmeister. In der Vorrunde in Bozen waren die Kanadier in Kneipen, Bars und Diskotheken gemessene Gäste - als es in Mailand um alles ging, blieben sie lammfromm. "Vor dem Finale haben meine Jungs nur Mineralwasser **getrunken** und waren früh im Bett", grinste Kingstone. Vor 33 Jahren holten die berühmten "Trail Smoke Eaters" zum letzten Mal den Titel nach Kanada - und ein Originaltrikot der "Rauchfresser" an der Kabinenwand erinnerte während des ganzen Turniers an das große Ziel. (094/MA1.43906 Neue Kronen-Zeitung, 10.05.1994, S. 55)

Flexion von *trinken* Rechtschreibung

Wortklasse: [Verb](#)
Stammformen: trinken / trank / getrunken
Hilfsverb: [haben](#)
Flexionsklasse: [unregelmäßige Verben](#)
Besonderheiten: [e-Tilgung im Konjunktiv II](#), [Ablaut in Stammformen](#)

Einfache Zeiten

Präsens			
Indikativ	Verb	Konjunktiv I	Verb
ich	trinke	ich	trinke
du	trinkst	du	trinkest
er/sie/es	trinkt	er/sie/es	trinke
wir	trinken	wir	trinken
ihr	trinkt	ihr	trinket
sie	trinken	sie	trinken

Präteritum			
Indikativ	Verb	Konjunktiv II	Verb
ich	trank	ich	tränke
du	trankst	du	tränkest
er/sie/es	trank	er/sie/es	tränke
wir	tranken	wir	tränken
ihr	trankt	ihr	tränket
sie	tranken	sie	tränken

Imperativ	
Person	Verb
Singular	trink
Plural	trinkt

Figure 4: An example of hyperlinked morphological information in the *Elexiko* German dictionary for the verb *trinken* ('to drink', left) pointing to its conjugation paradigm in the Canoo morphological lexicon (right).

in the primary-level vicinity of the headword (as with the Collins English dictionary for learners or the ANW scholarly dictionary of contemporary standard Dutch), while other languages place this information on a secondary level accessed by clicking on an additional button or tab (as with the DAELE Spanish Learners' Dictionary or the Great Dictionary of Polish illustrated in Figure 5).

The screenshot shows the entry for 'sadzonka' in the 'Wielki słownik języka polskiego'. The main entry is 'sadzonka' with a sub-entry '1. roślina'. Below it, the part of speech is 'rzeczownik' and the grammatical gender is 'ż'. The 'Odmiana' (Inflection) section lists various forms of the word:

liczba pojedyncza		liczba mnoga	
M:	sadzonka	M:	sadzonki
D:	sadzonki	D:	sadzonek
C:	sadzonce	C:	sadzonkom
B:	sadzonkę	B:	sadzonki
N:	sadzonką	N:	sadzonkami
Ms:	sadzonce	Ms:	sadzonkach
W:	sadzonko	W:	sadzonki

The right sidebar contains navigation options: Definicja, Kwalifikacja tematyczna, Relacje znaczeniowe, Połączenia, Cytaty, Odmiana, and Pochodzenie.

Figure 5: An example of embedded morphological information in the Great Dictionary of Polish for the noun *sadzonka* ('a seedling') in the *Odmiana* (Inflection) section of the dictionary entry.

Given that the Sloleks morphological lexicon is planned to include both standard and non-standard basic and inflected forms, labelled with the corresponding variation type and its compliance with the language norm (see chapter by Dobrovoljc et al. in this book for a detailed description), the morphological lexicon thus also functions as the pivotal source of information on potential spelling, pronunciation, inflection, derivation, syntactic or other issues related to individual lexical items. In addition to the lexicon providing the lists of all variant forms or pronunciations, their classification by specific variation type also allows for automatic selection and display of the relevant language issue explanation(s) in the norm-related section of the dictionary entry.⁸ Using the same mechanism, specific tags

⁸ The norm-related section of the dictionary, proposed by Krek et al. (2013: 41), is designed as a style guide with user-friendly explanations of language issues in Slovenian. The explanations are based on the ontology of most frequent types of linguistic issues in Slovenian (Krek and Dobrovoljc 2011), and are thus designed as a set of universal explanations to be displayed with all lexical items related to a certain type of issue.

or notifications can be automatically displayed in different parts of the dictionary entry (for example, next to the headword or one of its variant spellings; next to a particular word form or pronunciation etc.) to alert users about specific issues or particularities and direct them to the related explanations.

In addition to being the source of morphological, grammatical and normative information, the morphological lexicon also has an essential role in displaying other types of dictionary information. It enables searching by all possible forms and spellings and therefore allows users to form intuitive search queries without having to consider the lemmatization, part-of-speech categorization and spelling principles used in the dictionary headword selection, as is currently the case with the reference DSL2 dictionary and the Fran dictionary portal.⁹ In a similar way, a morphological lexicon can enhance the comprehensibility of definitions by linking individual word forms with the relevant lexical units (see for example the hyperlinking mechanism in the definitions of Wiktionary and TheFreeDictionary), or by linking the dictionary to external language resources and tools, as in the case of the Sloleks web service,¹⁰ where clicking on a particular word form or lemma takes the user to the list of all relevant concordances in the reference corpus (i.e. usages of the word form in context). Similarly, information on phonetic transcription in the background lexicon enables machine-generated speech synthesis of displayed word forms on the one hand, and automatic speech recognition of voice search queries on the other.

4.2 Relation between lexicon data and dictionary information

Despite the many technical and content-related advantages of keeping morphological information in a separate database, we must distinguish between original data in the lexicon database on the one side and the user-oriented dictionary information on the other, when planning its visualisation. One of the main advantages of a hierarchically-organized machine-readable system is the fact that it enables dynamic adjustments of information visualisation with respect to the type of language manual or the specific needs of its users. These include not only graphical design and technical solutions, but also the selection of the displayed information itself, such as the inclusion of data on non-standard language use, pronunciation or specific grammatical information, discussed below.

⁹ For example, the log file analysis of the Danish *Den Danske Netordbog* online dictionary (Bergenholtz and Johnsen 2005: 127–133) shows that the 19.5% of unsuccessful searches mostly include the passive and imperative forms of verbs, misspellings, spelling mistakes affected by pronunciation, and mistakes in writing multi-word expressions as one or several words.

¹⁰ <http://eng.slovenscina.eu/sloleks>

Most existing dictionaries of the Slovenian language include information on both standard and non-standard inflected forms. However, the latter are usually limited to a closed set of most common orthographical and morphological exceptions, such as the declension of nouns *otrok* 'child', *mati* 'mother', *hči* 'daughter', *gospa* 'lady/Mrs', and so on. A usage-based morphological lexicon, compiled to give an exhaustive description of formal characteristics of Slovenian lexica, would also include all frequent variant irregular patterns and modifications, such as the non-standard phoneme additions in declension, sound changes, etc. Experience in visualising the Sloleks morphological lexicon, which already includes several demo instances of such variant paradigms, shows that users prefer to see standard paradigms written in full, regardless of the frequency of usage of individual inflected forms, whereas the addition of full non-standard paradigms is too difficult to process, so the visualisation of these should be reduced to individual inflected forms occurring in corpus data. One of the first priorities of future user-experience research is thus to determine the frequency threshold, below which displaying non-standard language usage information no longer plays an informative or educational role, but instead acts as a disruption in the overall comprehensibility of the given information, regardless of the graphic design solutions.

A similar issue arises when visualising information on pronunciation, as the high frequency of stress placement variants in the Slovenian language results in extensive pronunciation paradigms; if we augment these by the alternative pronunciations of particular phonemes or different types of pronunciation transcriptions (accentuated or unaccentuated word forms; standardised or customized phonetic transcription), the display of all the combinatorics of all possible word forms quickly becomes overwhelming. It is thus important to prioritize pronunciation information according to its relevance to dictionary users, for example, show the accentuated headword and its phonetic transcription by default, but embed other phonetic information, such as the full accentuated inflectional paradigm or its phonetic transcription (only rarely found in general dictionaries), on a secondary level accessed in a separate section or a special extension button next to the default, unaccentuated inflectional paradigm.

The third important aspect to consider when distinguishing lexicon data from dictionary information is the visualisation of grammatical information. The formal grammar used in the compilation of a morphological database, usually adjusted to meet the needs and limitations of automatic natural language processing, is not necessarily equivalent to the grammar description given in a general dictionary. In addition to terminological considerations, such as renaming particular grammatical features that might be less comprehensible for non-linguistic users (e.g. non-definiteness or biaspectuality), and an evaluation of their actual relevance for the user, this also includes elemental linguistic decisions on the inventory of part-of-speech and other morphological categories, as well as the criteria for their

selection with particular lexical items.¹¹ Using different approaches for different types of lexical databases is not problematic in itself, but it is important that the specific mappings between the two are systematized and well-documented, as this is a prerequisite for the full compatibility of fundamental language resources, such as a morphological lexicon, a lexical database or a dictionary, and their long-term usability in other language resources and tools.

4.3 Relation between lexicon and dictionary entry

Since a morphological lexicon is primarily intended to store information on the inflectional, derivational, normative and other morphological properties of lexical items, and not their semantic characteristics, lexical items with identical morphological, phonological and grammatical features are usually merged into one lexicon unit, regardless of potential differences in meaning. In this way, the Sloleks lexicon merges homonymous semantically distinct lexical items, such as *bor* ('pine tree') and *bor* (the chemical element), or *početi* ('to start') and *početi* ('to do'), into a single lexicon entry, while semantically equivalent, but formally different lexical items, such as *volivec* and *volilec*, *posebej* and *posebaj*, *zvedeti* and *izvedeti* (two different spellings of 'a voter', 'especially' and 'to find out', respectively), are separated into two or more distinct lexicon entries. When integrating a morphological lexicon into a general dictionary or its underlying lexical database, it should thus be remembered that given the different designs and purposes of both databases the relation between the lexicon and dictionary entries is not necessarily symmetrical nor static, as it primarily depends on the consensually defined criteria on what constitutes the basic unit (an entry) in each database.

One of the key dictionary design decisions, influencing the way the lexicon and the dictionary database inter-connect, is undoubtedly the selection of formal criteria used for distinguishing homonymy from polysemy. That is, defining what formal properties of two semantically distinct lexical units with identical spellings should be considered when deciding whether to describe them in separate dictionary entries (homonymy), or within the same dictionary entry with two or more different meanings (polysemy). According to Gantar (2015: 341), both theoretical and user-orientated lexicographical approaches to the issue of homonymy-polysemy distinction usually agree that differences in one of the following formal characteristics should be considered as sufficient criteria for homonymy to be chosen, regardless of the degree of semantic or etymological similarity between the two items: homographs belonging to different

¹¹ An example of such differences in grammatical information in the morphological lexicon on the one side and a general dictionary on the other, would be potential merger of adverbial participles (currently stored as adverbs in Sloleks) with their original verbs, or elatives (currently stored as separate entries in Sloleks) with other degrees of comparison, etc.

part-of-speech categories (e.g. the noun and the adverb *naglas* ‘accent/loudly’, the noun and the adjective *žužkojed* ‘insectivore/insectivorous’); homographs with different grammatical features (e.g. the masculine and feminine noun *prst* ‘finger/soil’, the masculine and neutral noun *čelo* ‘cello/forehead’); homographs with a different inflection (e.g. the imperfective verbs *vesti-vezem* and *vesti-vedem* ‘to embroider/to behave’ or the perfective verbs *postati-postanem* and *postati-postojim* ‘to become/~to pause’); or homographs with different pronunciations (e.g. *molíti-molím* and *móliti-mólim* ‘~to hand out/to pray’ or *partíja-partíje* and *pártija: pártije* ‘a political party/a match’). This is also in line with the entry selection criteria used in the Sloleks morphological lexicon, which separates all these items into two distinct lexicon units – the relationship between the lexicon and the dictionary entry is thus symmetric.

However, there is less lexicographic consensus on whether the formal properties to be taken into consideration also include: the differences in derivation (e.g. the homonymous noun *vila* ‘a villa/a fairy’, where the derived adjective *vilinski* is only associated with the second of the two meanings); the differences in part of the inflectional paradigm (e.g. the homonymous adjective *bučen* ‘loud/of-pumpkin’, where the comparative forms are only associated with the first of two meanings, or the homonymous noun *lisica* ‘a fox/handcuffs’, where the second meaning is only associated with plural forms); or the differences in specific inflected forms (e.g. the homonymous noun *tenor* ‘the voice/the singer’, where the two items only differ in the singular accusative form that depends on animacy). Given that the Sloleks lexicon has also been designed to be used in natural language processing applications, which are not yet capable of reliable semantic disambiguation of identical inflected forms with identical grammatical features (e.g. disambiguating the form *bučnega*, *lisic* or *tenorja* in all different possible meanings), the lexicon thus follows the principle of the maximum possible paradigm that merges such overlapping inflectional paradigms into a single lexicon entry, even if specific meanings only take on a limited subset of all possible forms. Regardless of whether or not these meanings are separated into independent entries in the dictionary, the relationship between the lexicon entry and the dictionary entry is thus inherently asymmetric, since a particular dictionary headword or one of its meanings only correlates with a subset of a certain lexicon entry (e.g. *lisica*, *bučen*, and *tenor*). This potential asymmetry of interlinked database entries should thus be given special consideration when designing the technical and visualisation solutions for an online dictionary.¹²

If the previous paragraph discusses relating one lexicon entry to several dictionary entries or meanings, it is equally relevant to address the issue of relating one

12 One possible solution on how to display meaning-dependent morphological information, can be observed in the Great Dictionary of Polish, which considers all homographs as polysemous items, regardless of their diachronic connection, but requires the users to select the meaning of interest before displaying any additional information on grammatical properties or inflection.

dictionary entry to several lexicon entries. A typical example of this kind of database asymmetry are lexemes with variant spellings, e.g. *žiroračun* and *žiro račun* ‘a giro account’, *eventuelno* and *eventualno* ‘possibly’, *volivec* and *volilec* ‘a voter’. The Sloleks lexicon stores these as distinct lexicon entries, whereas a dictionary usually considers them to represent the same lexical item if no semantic differences are observed, merging their description into a single dictionary entry with several spellings and related inflection or pronunciation paradigms. A similar issue arises when describing (potentially) semantically identical pairs of homographs with variant grammatical lexical features, as with *činčila* ‘chinchilla’, *sluz* ‘slime’ or *nadlaket* ‘upper arm’, which are used both as masculine or feminine nouns, or with *finale*, which is used both as masculine or neutral noun, without any change in meaning. Even if a dictionary considered these to constitute separate dictionary entries (i.e. in symmetry with the lexicon), displaying information for several lexicon entries within a single dictionary entry is nevertheless inevitable for lexemes with multi-gender inflections, as for example the neutral noun *oko* ‘eye’ that takes the feminine plural form *oči* in one of its meanings, or the feminine noun *ledvica* ‘kidney’ that takes either the feminine (*ledvice*) or the neutral (*ledvica*) plural form.

5 CONCLUSION

Both lexicographic tradition and empirical user research confirm that morphological information represents an indispensable part of lexica description in a general dictionary. In order to meet this information need, it seems that future dictionaries of the Slovenian language should break with the tradition of presenting inflectional information in the abbreviation system that was created for the print-based design of the first edition of DSLL, given the limited degree of comprehensibility and the general technological advances that have taken place over the past few decades. With respect to the rich morphology of the Slovenian language, keeping this information in the form of a separate database brings many advantages to dictionary makers and users alike. However, its integration into a dictionary must be designed and implemented in a systematic way, so as to ensure the dictionary’s long-term compatibility with other language resources and tools, and to enable its dynamic adjustment to meet the varying needs of diverse user groups.