Online Style Guide for Slovene as a Language Resources Hub

Simon Krek[1], Helena Dobrovoljc[2], Kaja Dobrovoljc[3], Damjan Popič[4]

[1]Jožef Stefan Institute, Ljubljana, Slovenia
[2]Fran Ramovš Institute for the Slovene Language, ZRC SAZU, Ljubljana, Slovenia
[3]Trojina, Institute for Applied Slovene Studies, Ljubljana, Slovenia
[4]Faculty of Arts, University of Ljubljana, Ljubljana, Slovenia
E-mail: simon.krek@ijs.si, helena.dobrovoljc@zrc-sazu.si,
kaja.dobrovoljc@trojina.si, damjan.popic@ff.uni-lj.si

Abstract

There exists a long tradition of orthography guides or style manuals for Slovene dedicated to "good writing" (Slo. *pravopis*, Ger. *Rechtschreibung*), with the first one published in 1899 and the most recent in 2001. The new web portal developed within the Communication in Slovene project is taking the concept originating from the world of print one step further into the digital environment, with a question-answering system which analyses the question entered into a query window in natural language and aims to provide a three-layered answer, from a more condensed and graphical one using data from extensive corpora, lexicons, dictionaries and other online resources, to a more general user-friendly description of the problem, together with links to digitized modern and historical normative resources related to the identified language problem. The paper describes a demo version of the portal with demonstration data for 15 language problems.

Keywords: Slovene language; orthography; online style guide; language resources portal; question-answering;

1. Introduction

The basic idea of the portal[1] is to provide information about the Slovene language and the problems that average speakers have with its written norm. It is not intended only for language specialists or professionals but for all web users. The portal uses new (language) technologies now available also for Slovene and aims to complement printed orthography guides from Levec (1899) to Toporišič (2001) with a dynamic web portal based on empirical data from various extensive language resources. The concept is based on the analysis of language use in text corpora and frequent questions in web forums dedicated to language problems, at the same time also providing information from traditional orthography guides and other historical resources. The most important extensive new digital language resources used on the portal are Sloleks morphological lexicon (Grčar et al., 2013) and Gigafida corpus (Logar Berginc et al., 2012).

[1] http://slogovni.slovenscina.eu/

2. Background

Similar to other languages (Mønnesland 1998: 1103) Slovene has a relatively long tradition of written language codification embodied in official orthography guides in the entire 20th century. These guides have usually included an extensive dictionary section, with an emphasis on orthographically challenging vocabulary (cf. Verovnik 2004: 254). The last orthography guide in the series was published in 2001 in printed form, on CD-ROM in 2003, and has been available online since 2010. The content of the digital version replicates the printed one, the rules are available as a PDF document, and dictionary content can be searched in the search engine NEVA,[2] on the Termania dictionary portal,[3] and in ASPplus software,[4] all of them also allowing more complex queries.

One of the assumptions of the authors of the new portal is that the advent of the web, with the possibility of massive participation of users in the creation of texts (blogs, forums, social networks, etc.) that are immediately available to be read or commented on, radically changed the nature and dynamics of the text publication process. In post WWII Slovenia, this process has typically included the author, the publishing house with its editor, the proof-reader, and a language specialist called "lektor" responsible for the compatibility of published texts with the language norm or standard.

In the world of print, texts have traditionally been handled by a relatively narrow circle of professionals, including language specialists. However, with the possibility to publish texts online without the assumed or axiomatic interference of third parties, this cycle is now more or less broken. In addition, the time needed from the creation of the text to its publication has been reduced to just a few seconds, and numerous genres previously reserved for private communication are now part of the public sphere (Crystal, 2011). This has created the need to also present information about language standard to the general public, not just language professionals, preferably in a user friendly manner. Therefore, if previous orthography guides effectively belonged to the world of print, the new web portal aims to provide an answer to the question of how language codification should be presented in the digital (web) environment of the 21st century.

In the new environment, codification-related language help currently comes from two basic sources. The first one comprises spelling or grammar checkers and similar tools which can be seen to replace the proof-reader in the printed environment. The other sources are online portals, dedicated forums and now also social networks, or

[2] http://bos.zrc-sazu.si/sp2001.html/

[3] http://www.termania.net/slovarji/20/slovenski-pravopis/

[4] http://www.amebis.si/aspplus/

search engines, providing consultation or feedback from both peer communities and official bodies responsible for language codification. The new web portal aims to answer the need for consultation by providing standardized explanations of the most frequent problems with language or (more narrowly) spelling and orthography.

3. List of language problems

The portal consists of several parts, with a list of around 700 detected language problems functioning as the central database. The list was created by analyzing traditional orthography guides, text corpora and web forums specialized in language problems. Web forums were crawled and each question was manually assigned to a particular category. Also, special data mining procedures were established which produced lists of variant forms of words where speakers (or writers) of Slovene falter due to inappropriate, unrecognized or non-existent norms. The main task in this process was to establish a list of real language problems and balance it suitably between overgeneralization and excessive fragmentation of categories. All categories were later organized as an ontology with eight top categories: orthography (A), orthoepy (B), morphology (C), word-formation (D), vocabulary (E), syntax (F), text (G), and other (H). Current ontology extends to six levels from top to bottom, with variable granularity. Levels are formally labelled as combinations of letters and digits, as shown in Table 1.

LABEL	CATEGORY
D	word-formation
D1	adjectives
D1a	possessive adjectives from names of masculine gender
D1a1	from names ending in vowels
D1a1a	from names ending in -a
D1a1b	from names ending in unpronounced -e
D1a1c	from names ending in -y

Table 1: An example of language problems ontology

4. Three-layered configuration of answers

Each of the bottom-level categories in the ontology is linked to several elements in the database, with the “short” and “long” answers (see Figure 1) the most important ones.

4.1 Short answer

The short answer consists of text in XML format which can generate a formulaic textual answer with relevant statistical data from the corpus and the lexicon. It is designed as a universal mechanism for the (statistical) description of all possible combinations of standard and non-standard word forms belonging to one particular category. For further clarification, category D1a2e will be used as an example:

LABEL	CATEGORY
D	word-formation
D1	adjectives
D1a	possessive adjectives from names of masculine gender
D1a2	from names ending in consonants
D1a2e	from names ending in pronounced -r

Table 2: Example – category D1a2e

The full title of the D1a2e category is "Word-formation of possessive adjectives derived from names of masculine gender ending in pronounced –r". Examples of (foreign) surnames in Slovene belonging to the category are Shakespeare, Baudelaire, etc. Most of the adjectives derived from these names have two variant forms with alternative endings *-jev* and *-ov*: *Shakespearjev* | *Shakespearov*, *Baudelairov* | *Baudelairjev*. Since the final unpronounced -e has to be dropped in the derivation process according to the standard, essentially changing the exact form of the original name, two non-standard forms are used frequently enough to be in included in the lexicon: *Shakespear**e**jev* | *Shakespear**e**ov*, *Baudelair**e**jev* | *Baudelair**e**ov*. Therefore, there are four potential forms that have to be taken into account when creating the short answer for this category. As it is not necessary that all four forms actually appear in the corpus for all possible names in this category, a combination of 15 answers have to be included in the short answer. Table 3 shows the first four:

<!-- variant 1: FOUR, standard-12, non-standard-34 --> <text var="S00.S00.N00.N00" graph="1234">The graph shows the data about the use of word forms <word id="1"/>, <word id="2"/>, <word id="3"/> and <word id="4"/> in the Gigafida corpus. Word forms in blue colour are standard, those in grey are not compatible with the current standard of written Slovene.</text>
<!-- variant 2: THREE, standard-12, non-standard-3 --> <text var="S00.S00.N00" graph="123">The graph shows the data about the use of word forms <word id="1"/>, <word id="2"/> and <word id="3"/> in the Gigafida corpus. Word forms in blue colour are standard, the grey one is not compatible with the current standard of written Slovene.</text>
<!-- variant 3: THREE, standard-12, non-standard-4 --> <text var="S00.S00.N00" graph="124">The graph shows the data about the use of word forms <word id="1"/>, <word id="2"/> and <word id="4"/> in the Gigafida corpus. Word forms in blue colour are standard, the grey one is not compatible with the current standard of written Slovene.</text>
<!-- variant 4: THREE, standard-1, non-standard-34 --> <text var="S00.N00.N00" graph="134">The graph shows the data about the use of word forms <word id="1"/>, <word id="3"/> and <word id="4"/> in the Gigafida corpus. The word form in blue colour is standard, those in grey are not compatible with the current standard of written Slovene.</text>

Table 3: Short answer in XML

Word forms shown in the textual part of the short answer (as opposed to the graph) are taken from the Sloleks lexicon which also contains statistical data from the Gigafida corpus. In each particular case, the system choses the relevant short answer automatically in accordance with the lexicon data. The design of short answers therefore enables an upgrade of the corpus which is directly reflected on the portal through the upgrade of the data in the lexicon. Once the set of possible short answers is written for a particular language problem, it is not necessary to update the text of the answer again manually, as the system choses the right answer according to the status found in the regularly updated lexicon. This makes the system dynamic and linked to external independent resources, which can be updated regularly. Where this is applicable, data from the lexicon/corpus are also shown in a graph. For visualization of the data, the portal uses Google Charts tools, as shown in the upper part of Figure 1.

4.2 Long answer

In contrast to short answers, which constitute the dynamic part of the portal linked to external resources, long answers are essentially static. Each identified problem in the ontology receives one long answer which is written in HTML format and included in the central database. When creating the system, special attention was given to wording, length, formatting and other features, to ensure that long answers are particularly useful for general users, who are the primary target audience of the portal, rather than language professionals.

Long answers (the middle part of Figure 1) can contain three types of links each with a different function:

- blue, italic, bold: link to an external resource, which can be a corpus, lexicon or other web resource such as Wikipedia, etc.
- blue, underline: pop-up window with an explanation of a linguistic term when its use is unavoidable in the long answer.
- blue, dotted underline: pop-up window with the list of words belonging to the same category, with the same orthographic problem.

Long answers are designed to provide the user with general information about the problem in lay terms, and contain links to other available resources that we consider useful for the user. This part of the portal has an explicitly educational function, as it is expected for the user to understand the problem and be able to interpret in the future.

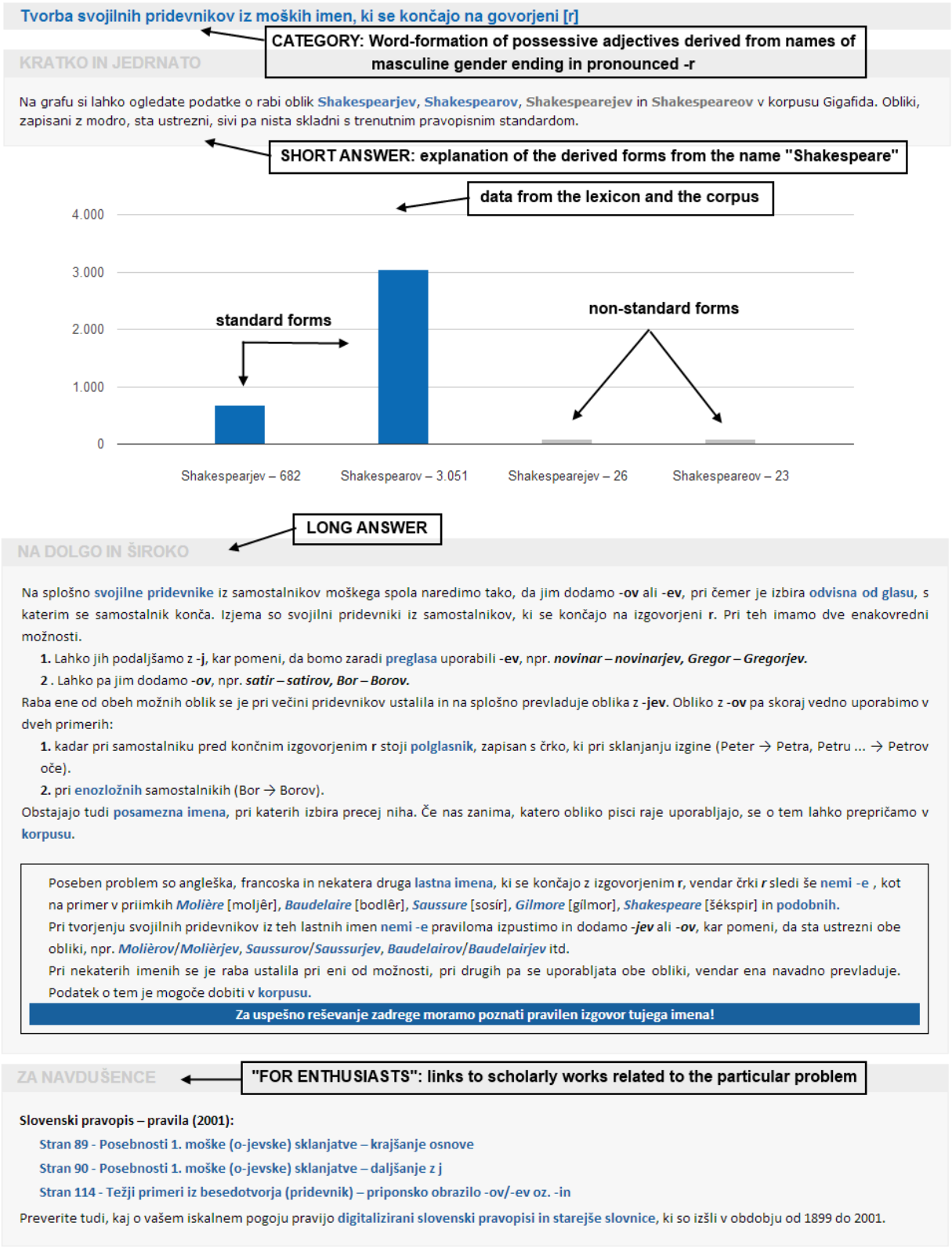

Figure 1: Screenshot of the query result on the portal

4.3 Links for enthusiasts

The third part of the answer is titled “For enthusiasts” and provides links to scholarly works related to the particular problem or to orthographic problems in general. The most important document in this section is the official orthographic rules book published in 2001 and available online in PDF format. Other important works include previous orthographic guides which were digitized in another project and published online independently,[5] and are also included on the portal. This part of the portal provides more advanced users with the possibility to explore the historical background of the problem encountered.

5. Access to information on the portal

Information on the portal can be accessed in two ways: first, by entering a query in natural language which is parsed and matched with the data in the lexicon. Parsing is performed by a rule-based tagger and parser owned by the Amebis software company.[6] Individual word forms and lemmas from the query are compared with lexicon entries that contain information about a category from the ontology of language problems. If a match is found, the corresponding answer is shown on the portal. If there is more than one match, other possibilities are shown as links in the “Did you mean?” section on the left side of the main frame. As some problems in the ontology are related to each other by default, if one is found, the others are shown in the “Linked answers” section.

The second option for accessing information is to browse the ontology on the index page which can be accessed by clicking the “See the index” link on the home page. Users who wish to go through the entire portal systematically can use this feature.

6. The corpus and the lexicon

The most important relationship, enabling the system to work as designed, is that between the ontology—with its formal hierarchy of labelled language problems—and the Sloleks lexicon containing extensive amounts of data about morphology, together with information about language norm assigned to its various elements. Gigafida corpus, on the other hand, as the source of statistical data for the lexicon, does not contain normative information. It is lemmatized and POS-tagged in a standard manner using the newly-developed Obeliks tagger and lemmatizer (Grčar et al., 2012).

[5] Available at: http://www.trojina.org/pravopisi/

[6] Web site: http://www.amebis.si/

The lexicon uses Lexical Markup Framework (LMF) format which allows various kinds of information to be included on every level, either assigned to the whole lexical entry or to one particular word form. These types of information can range from pronunciation or stress to normative information. One particular instance of the lexicon, i.e. lexical entry, becomes a part of the portal only when it is assigned with a particular language problem from the ontology. Without explicit information it is invisible to the system. The annotation of normative information in the lexicon is currently performed semi-automatically or manually, as this kind of information is too sensitive to be included in a fully automatic manner without checking.

6.1 Extraction of data from the corpus

In order to obtain a candidate list of lexicon entries for a particular language problem, an extraction procedure is applied to the corpus. To explain the procedure in detail, category C1a3b will be used: "Declension of (foreign) names of masculine gender with the 'unsteady vowel'". Examples of such names in Slovene are Russell, Powell or Robben, Bremen which lose their final [e] in some grammatical cases: *Russlla*, *Powlla* or *Robbna*, *Bremna*. Since this rule can produce rather unusual forms with a series of consonants, Slovene writers often use the final [e] in inflected forms: *Russ**e**lla*, *Pow**e**lla* or *Robb**e**na*, *Brem**e**na*.

To extract relevant names from the corpus, in order to decide which names will be later included in the lexicon, all types in the corpus are split into three parts: the root (open set), the inflections (closed set) and the variable part (closed set). Based on the variability of the middle part and the invariability of the other two, pairs of types are produced, together with frequency data. The more equally the variable part is distributed between both possible forms, the more interesting the pair. When the extracted pairs are ranked according to the combination of frequency and variability using statistical data from the corpus, a list shown in Table 4 is produced. As this category covers different combinations of an 'unsteady vowel' + a consonant (en/-n-, -ek/-k-, -ic/-c-, -ell/-ll-, etc.), for each consonant pair a separate list is prepared. Table 4 shows the top 20 candidates for the **en/-n-** pair. These traditionally include names of Scandinavian or Germanic origin which is also confirmed on the extracted and ranked list.

Extraction of corpus data enables the portal to offer information about the most challenging and frequent names belonging to this category, and on the other hand, long-lived examples from traditional resources can be replaced with modern and relevant ones in long answers.

Root	**Lemma (artificial)**	**Frequency in Gigafida**		**Score**
		root + **-en-** + inflection	root + **-n-** + inflection	
Klem	Klemen	1843	3839	0,46
Lor	Loren	908	505	0,29
Berg	Bergen	208	375	0,25
Niels	Nielsen	164	120	0,25
Test	Testen	501	2326	0,24
Robb	Robben	163	333	0,24
Natlač	Natlačen	223	147	0,23
Gold	Golden	37	29	0,21
Gall	Gallen	105	148	0,20
Ols	Olsen	112	64	0,20
Bid	Biden	102	117	0,20
Bjorndal	Bjorndalen	112	163	0,20
Franz	Franzen	117	114	0,19
Jens	Jensen	138	60	0,19
Patt	Patten	85	113	0,19
Hag	Hagen	74	120	0,19
Brem	Bremen	220	1509	0,18
Hold	Holden	60	147	0,18
Jem	Jemen	196	1319	0,18
Bed	Beden	769	164	0,18
Dresd	Dresden	194	1410	0,18

Table 4: Names extracted from the corpus and ranked according to frequency and variability

6.2 Manual analysis of corpus data

In some cases, extracted lists do not need further analysis and can be used for lexicon upgrade immediately. However, in most cases they are treated as candidate lists which have to be checked manually, either to validate data (corpus noise) or because different variants have to be attributed with unpredictable normative labels. For this purpose, the crowdsourcing platform sloCrowd (Tavčar et al., 2012) is used. The system supports annotator authentication and supervision, as well as quality control through random check based on gold-standard data. To explain the procedure in more detail we will use category C1a3f (Table 5):

LABEL	CATEGORY
C	morphology
C1	nouns
C1a	nouns of masculine gender
C1a3	nouns of masculine gender ending in vowels
C1a3f	names ending in -y

Table 5: Example using category C1a3f

This category is dedicated to (foreign) names ending in written [y] pronounced either as /ɪ/ or /e/, or a diphthong /ɑɪ/, /ɔɪ/, etc., such as Harry, Sydney, Playboy, Orsey, etc. In the Slovene declension system, these nouns are treated differently if they are pronounced with the final single vowel or a diphthong. In the first case, standard inflections are extended with a -j- before the inflection while in the second case this is not needed since the diphthong itself is considered to contain the sound /j/ in Slovene. Therefore, the examples mentioned above have the following forms in genitive case singular: *Harry**ja***, *Sydney**ja***, *Playboy**a***, *Orsey**ja***. *Playboy* is pronounced with a final diphthong and has a regular inflection; others have to be extended with the medial -j.

The initial extracted list contains all names with the final written y. However, those with the consonant + y combination can be excluded from manual analysis as their pronunciation is predictable, and therefore both standard and non-standard inflectional paradigms are predictable and can be included in the lexicon automatically. With names ending in the vowel + y combination pronunciation is not predictable and manual procedure is needed to determine first the standard pronunciation of the foreign name, and based on that, the standard or non-standard inflectional paradigms.

For this purpose, a task is defined in the sloCrowd software, as shown in Figure 2, and results are obtained based on three or five decisions depending on the difficulty of the task. In the pilot project, around 100 students from the Faculty of Arts (Department of Translation) at the University of Ljubljana worked on approximately 8,000 extracted names in 10 tasks.

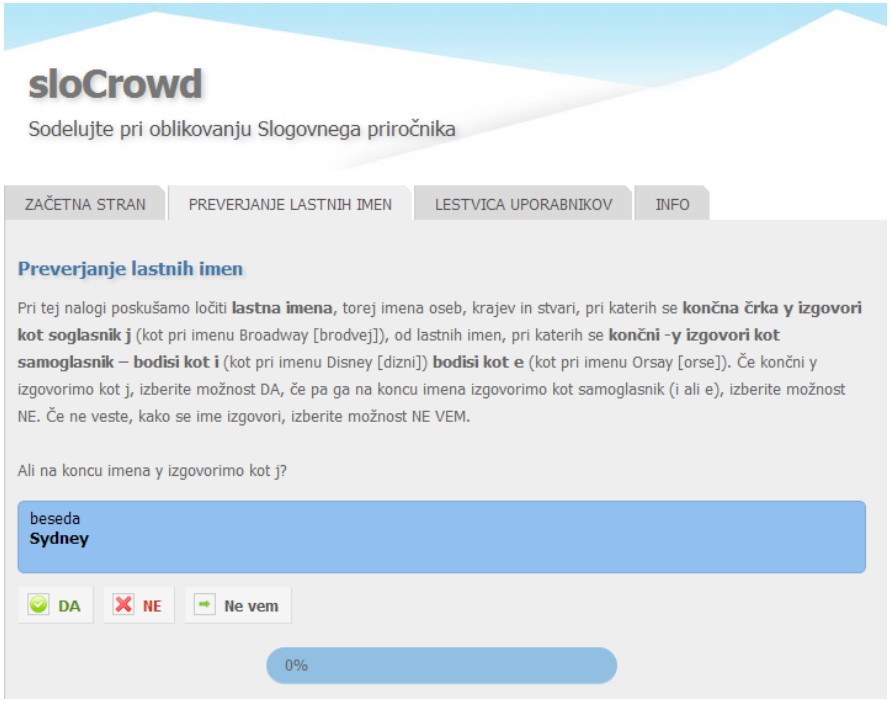

Figure 2: Screenshot of a task in the sloCrowd crowdsourcing software

6.3 The lexicon

Sloleks lexicon is an independent language resource in the LMF (XML) format and can be found at different web addresses, both for downloading and for searching.[7] Elements from the lexicon become part of the portal if they contain information about a category from the ontology of language problems (attribute "SPSP"), normative labels (attribute "norma") and norm types (attribute "tip"). This additional information is added to the standard information which includes the description of formal morphological features of lemmas and word forms: morphosyntactic descriptions or MSDs.

Attribute "norma" (=norm) can have three values: *non-standard*, *variant* or *unclear*. *Variant* is used when several alternative forms can be used according to the standard, and *unclear* is used when the normative status of a lemma or word form cannot be determined due to conflicting information in the rules and dictionary parts of the official orthography guide. The absence of the attribute signifies that the lemma or word form is standard.

Attribute "tip" is used for differentiating between two or more possible morphological paradigms within one lexical entry, and related to one category, as shown in the example from lexicon in Figure 3. The lemma denotes the Slovene masculine name "Matija" which has two legitimate inflectional paradigms; therefore, the value in the attribute "norma" is *variant*. The two possible forms for genitive singular (=morphosyntactic description *Slmer* in the "msd" attribute) are *Matija* and *Matije*. The first paradigm is differentiated from the other using the attribute "tip" with the value which includes the category label, "s" for "standard form" and a sequential number for each paradigm.

Lexicon as a resource linking the portal and the corpus is used primarily for top level categories *orthography*, *word-formation*, *morphology* and *orthoepy*, and less commonly for *syntax*, *vocabulary* and *text*. For the latter three categories, data are generated either directly from the corpus or are not required, as answers are general enough to be limited to the long answer itself without the need for more detailed explanations.

[7] Download at: http://www.slovenscina.eu/sloleks/prenos or search: http://www.slovenscina.eu/sloleks.

```
<LexicalEntry id="LE_S_Matija" xmlns:d="urn:LEKSIKON_SSJ">
      <feat att="besedna_vrsta" val="samostalnik" />
      <feat att="vrsta" val="lastno_ime" />
      <feat att="spol" val="moški" />
      <feat att="SPSP" val="C1a2a" />
      <Lemma>
            feat att="zapis_oblike" val="Matija" />
      </Lemma>
<…>
      <WordForm>
      <feat att="število" val="ednina" />
      <feat att="sklon" val="rodilnik" />
            <FormRepresentation>
                  <feat att="zapis_oblike" val="Matija" />
                  <feat att="msd" val="Slmer" />
                  <feat att="SPSP" val="C1a2a" />
                  <feat att="norma" val="variantno" />
                  <feat att="tip" val="C1a2a_s_1" />
                  <feat att="pogostnost" val="858" />
            </FormRepresentation>
            <FormRepresentation>
                  <feat att="zapis_oblike" val="Matije" />
                  <feat att="msd" val="Slmer" />
                  <feat att="SPSP" val="C1a2a" />
                  <feat att="norma" val="variantno" />
                  <feat att="tip" val="C1a2a_s_2" />
                  <feat att="pogostnost" val="4018" />
            </FormRepresentation>
      </WordForm>
<…>
</LexicalEntry>
```

Figure 3: Sample from the lexicon in Lexical Markup Framework format

7. Conclusion

This article describes a new web portal dedicated to problems with Slovene orthography, and includes in its demonstration version data for 15 language problems in Slovene selected from the approximately 700 problems identified by analysing traditional reference books, web forums and different extensive text corpora. The portal uses two resources to present information about real modern Slovene to the users of the portal in a user-friendly manner: the 1.2 billion-word corpus Gigafida, and the Sloleks morphological lexicon with 100,000 lemmas, together with their inflectional paradigms.

The portal is built around a central database with the 700 language problems organized in an ontology with eight top-level categories. These categories are used to identify relevant parts of the lexicon with normative information, which enables the

system to use both lexicon and corpus data to present normative information on the portal in a standardized manner. This comprises three types of answers: the short answer with statistical data, also supplied in graphical form; the static long answer for each of the bottom-level categories; and links to scholarly books and documents for experts and enthusiasts. The article describes both the portal and the extraction of relevant word forms and lemmas from the corpus, which are later assigned with normative labels and included in the lexicon, also using crowdsourcing in the process.

8. Acknowledgements

This article is based on the work of the Communication in Slovene project, which is part-financed by the European Union, the European Social Fund, and the Ministry of Education, Science and Sport of the Republic of Slovenia. The operation is being carried out within the operational programme Human Resources Development for the period 2007–2013, developmental priorities: improvement of the quality and efficiency of educational and training systems 2007–2013.

9. References

Crystal, D. (2011). *Internet linguistics: a student guide*. New York, Routledge.

Grčar, M., Krek, S., Dobrovoljc, K., (2012) Obeliks: statistical morphosyntactic tagger and lemmatizer for Slovene. *Proceedings of the Eighth Language Technologies Conference, October 8th-12th, 2012*. Institut Jožef Stefan, Ljubljana, pp. 42-47.

Levec, F. (1899). *Slovenski pravopis*. Na Dunaju: cesarska kraljeva zaloga šolskih knjig.

Logar Berginc, N., Grčar, M., Brakus, M., Erjavec, T., Arhar Holdt, Š., Krek, S. (2012). *Korpusi slovenskega jezika Gigafida, KRES, ccGigafida in ccKRES: gradnja, vsebina, uporaba*. Trojina, Ljubljana.

Mønnesland, S. (1998). Emerging Literary Standards and Nationalism. The Disintegration of Serbo-Croatian. *Actas do I Simposio Internacional sobre o Bilingüismo*. 1103–1113.

Tavčar, D. Erjavec, T., Fišer, D. (2012). sloWCrowd: orodje za popravljanje wordneta z izkoriščanjem moči množic. *Proceedings of the Eighth Language Technologies Conference, October 8th-12th, 2012*. Institut Jožef Stefan, Ljubljana, pp. 197-202.

Toporišič, J. (2001). *Slovenski pravopis*. Založba ZRC, Ljubljana.

Verovnik, T. (2004). Norma knjižne slovenščine med kodifikacijo in jezikovno rabo v obdobju 1950–2001. *Družboslovne razprave XX*. 241–258.