Slovenski jezik – Slovene Linguistic Studies 12 (2019): 3–19

Alenka Jelovšek, Tomaž Erjavec
ZRC SAZU, Fran Ramovš Institute of the Slovenian Language, Ljubljana, Jožef Stefan Institute, Ljubljana

A corpus-based study of 16th-century Slovene clitics and clitic-like elements

The article presents a corpus-based linguistic investigation of the spelling variation in 16th century Slovene, focused on clitics and clitic-like elements. Based on a sample from three works of Slovene Protestant writers that was automatically modernised and then hand-corrected, cases where one original word is represented as two modernised words and vice-versa are analyzed, as well as synchronic orthographical variants of the most commonly bound clitics and their distribution, thus contributing to the general description of the spelling variation of the period.

Članek predstavlja na korpusu osnovano jezikoslovno raziskavo pisne variantnosti v slovenskem knjižnem jeziku 16. stoletja s poudarkom na naslonkah in njim podobnih elementih. Na podlagi vzorca iz treh del slovenskih protestantskih piscev, ki je bil avtomatsko posodobljeni in nato ročno popravljen, so analizirani primeri, ko eni izvorni besedi ustrezata dve posodobljeni in obratno. Predstavljene so tudi pisne različice najpogostejših naslonk, ki so v 16. stoletju zapisane skupaj z naglašeno besedo, in njihova distribucija, s čimer se dopolnjuje opis pisne variantnosti v obravnavanem obdobju.

1 Introduction

When the first standard written version of Slovene was established in the second half of the 16th century, the texts produced during this period displayed pioneering attempts at establishing an orthographic paradigm, as written Slovene had been used only sporadically between the end of the 10th and the middle of the 16th century. Lacking an established spelling tradition of their language, Slovene Protestant writers resorted to adapting the New Latin and Early New High German orthographic systems in various ways to represent Slovene (Slavic) phonemes, causing discrepancies and clashes within and across different texts (cf. Neweklowsky 1984: 394, Ahačič 2014: 262–267). These and other examples of spelling variation – combined with a considerable number of

inconsistencies – contributed to creating a high degree of variation at the onset of written Slovene in the early modern period.

While the topic has been dealt with in numerous studies using traditional linguistic methods (e.g. Merše, Jakopin, Novak 1993; Toporišič 2008/2009), this paper undertakes a corpus-based linguistic investigation of the spelling variation in 16th century Slovene both from the diachronic and synchronic points of view. The investigation is based on a manually annotated sample from three books, and it concentrates on clitics and clitic-like elements. We identify such words by comparing the spelling conventions of the early modern period to those of contemporary Slovene using normalised forms of the originals, where we observe cases where one orthographic word is nowadays written as two or more words (1–n mapping) or vice-versa (n–1 mapping). The normalisation also enables the identification of the orthographical variants of the most commonly bound clitics and their distribution, thus contributing to the general description of the spelling variation of the period.

This study builds on the work presented in Erjavec and Jelovšek (2013) and presents a first attempt to analyse clitic-like elements in 16th century Slovene, as these elements stand out as some of the most widely-occurring, problematic cases, which warrant analytical methods beyond a manual, traditional approach. It also provides an example of how corpus-based methods can be used to more thoroughly explore spelling variation based on a large quantity of data.

The rest of the paper is structured as follows: Section 2 details the data samples that are used in our analysis, Section 3 gives the linguistic analysis of the data, and Section 4 gives some conclusions and directions for further work.

2 The data

In our investigation we have concentrated on samples from three important 16th century Slovene books, given below with their identifiers, which we will use in the rest of this paper:

- TT 1557: Primož Trubar. *Ta pervi deil tiga Noviga teſtamenta* [The First Part of the New Testament] (Tübingen, 1557);
- JPo 1578: Jurij Juričič. *Poſtilla* [Postil] (Ljubljana, 1578);
- TPo 1595: Primož Trubar. *Hiſhna poſtilla* [House Postil] (Tübingen, 1595)

The books were chosen for their high degree of orthographical variation. As Primož Trubar was the author of the first Slovene printed book (*Cathechismus*, 1550), thus establishing basic orthographical rules of the period, two of his works were included in the analysis: the first one from the period when Trubar was still

building his orthographical system and the other his last, posthumously published work that was edited by his successors and highly influenced by their often different spelling choices. The third work included in the analysis, Juričič's *Postil*, was chosen for the same reason. Jurij Juričič was historically much less significant author, but being an immigrant from Croatia, he was influenced by his native tongue and its orthographical tradition (cf. Rigler 1968: 194–195, 221–222).

The three books were already available in a digital diplomatic transcription[1], i.e. one that was faithful to the printed books. The transcriptions were encoded in XML with mark-up for various structural units, as well marking the language of a span of text. From this basis we first constructed a corpus containing all the running Slovene texts of the three books, i.e. without the front pages, page numbers, printer's marks, or German text. This corpus was then automatically tokenised, i.e. split into words and punctuation symbols, and segmented to sentences. As the texts are quite long, this complete corpus includes almost 30,000 sentences and over 865,000 words, with TT 1557 having about 163,000, JPo 1578 215,000 and TPo 1595 487,000 words.

We then made a sample from each book, initially encompassing 200 sentences, which were first automatically modernised with the CSMTiser tool[2] (Ljubešić et al. 2016; Scherrer and Ljubešić 2016), which is a machine learning program that was trained on the manually word-modernised goo300k corpus of historical Slovene (Erjavec 2015). The automatic modernisation was then hand-corrected using the WebAnno platform (Eckart de Castilho et al. 2014). In the annotation process, sentences that were very short or contained errors were removed.

Book	Sentences	Tokens	Words	Original types	Modern types	Modern / Original
TT 1557	64	2,375	2,009	876	830	94.7%
JPo 1578	112	5,446	4,563	2,013	1,851	92.0%
TPo 1595	141	8,417	7,125	2,248	2,024	90.0%
ALL	317	16,238	13,697	4,471	3,547	79.3%

Table 1. Statistics over the samples.

Table 1 gives the final counts over the samples, split by book and in total. Together, the dataset contains just over 300 sentences and 16,000 tokens or almost 14,000 word tokens, the other tokens being punctuation symbols. The size of the three samples is considerably different, as different numbers of sentences were

[1] The digital library that contains also the complete three books is available at https://stage.termania.net/korpus16/.

[2] The tool is available from https://github.com/clarinsi/csmtiser.

deleted from the initial samples. In the table, the "types" columns give the number of different (lower-cased) original and modernised words, i.e. the lexicon of each sample and their union. The original types correspond to the words in the transcriptions, and the modern ones to the manually normalised ones; as could be expected, the latter is lower than the former, as normalisation removes variation in spelling. The last column gives the ratio between the two, and shows that, interestingly, there is least book-internal spelling variation in the first and oldest book (about 5%), while the second has about 8%, and the third as high as 10%, which can be attributed to the influence and inconsistency of the editors of the text (cf. Rigler 1968: 203–204, 230). Finally, the last row shows that, expectedly, spelling variation taken over all three book samples together is much greater, with almost 20% reduction in the size of the modern word-form lexicon.

Book	Words	n-m	%	1-n	%	n-1	%	Types	n-m	%	1-n	%	n-1	%
TT 1557	2,009	116	5.8	94	4.7	24	1.2	876	78	8.9	62	7.1	18	2.1
JPo 1578	4,563	387	8.5	362	7.9	28	0.6	2013	264	13.1	240	11.9	27	1.3
TPo 1595	7,125	272	3.8	208	2.9	85	1.2	2248	180	8.0	134	6.0	51	2.3
ALL	13,697	775	5.7	664	4.8	137	1.0	4471	493	11.0	411	9.2	92	2.1

Table 2. Statistics over clitic-like elements in the samples.

Because the focus of the paper is on clitics and clitic-like elements, we quantify the samples in terms of how many words have changed their word boundaries. Table 2 first repeats the number of word tokens and then gives the number of cases of word-boundary changing words, also divided into 1–n (splits) and n–1 (joins), also with percentages over the total number of word tokens. After the tokens, this information is also given for the lexicon of the samples and overall.

As can be seen, the overall percentage of split and joined word tokens is 5.7%, with JPo 1578 having the highest percentage, and TPo 1595 the lowest, less than half of that of JPo 1578.[3] Of these, the vast majority is for cases where a word is now split, with merges constituting, overall, only about one sixth of the cases. Interestingly, while JPo 1578 has the highest ratio of splits, it has the lowest number of joins. In terms of the lexicon, the proportion of split or joined words is about twice as high as for tokens, meaning that they tend to be fairly low frequency words, and this holds for both splits and joins.

[3] The difference can be mainly attributed to the spelling of non-syllable prepositions in TPo 1595 with an apostrophe and considered freestanding (see below).

3 Analysis

The sample corpus[4] was first imported into the noSketch Engine concordancer of the CLARIN.SI research infrastructure[5] and the analysis was performed using this tool. The noSketch Engine offers powerful functionality for corpus analysis:[6] it supports annotated corpora and the so called CQL language for querying, allows displaying of concordances as well as wordlist, etc. So, for example, it is simple to restrict the query only to parts of the corpus (such as samples from only one book), or to query via regular expressions (such as "k.*" for all words beginning with "k".

In the import the split and merged tokens were joined with the underscore, "_", with the original tokens having the default "word" attribute, and the normalised ones the "norm" attribute. It is thus easy to find all the clitic tokens, in particular, these are the CQL queries that return all the concordances of the such tokens:

- *[word=".*_.*"]* i.e. find all word tokens that contain any string, followed by the underscore and followed by any string,
- *[norm=".*_.*"]* i.e. the same but over the "norm" attribute.

The normalisation also enabled the identification of the spelling variants of the most commonly bound clitics – non-syllable prepositions *k* 'to', *v* 'in' and *z* 'with' – in the analysed sample. As the analysis of the preposition *v* showed the influence of traditional non-distinguishing between the consonant /v/ and the vowel /u/, a parallel relationship between the consonant /j/ and the vowel /i/ was also examined even though it rarely occurred in the clitics. The results are presented below.

3.1 Two modernised words as one original word

As shown in Table 3, the most predominant among the bound words are non-syllable prepositions *v* 'in(to)', *k* 'to', and *z* 'with' (cf. Novak 2011: 127) that account for 75% of instances in the analysed corpus while also displaying a high

[4] Although not further discussed here, the complete corpus is composed of two subcorpora: the one we are discussing here, named "clt", and another one, named "rnd", which was also manually annotated and is roughly the same size as the "clt" one. The two subcorpora are in the corpus distinguished by the value of the "subset" attribute on paragraphs, so "<p subset="clt">" identifies the paragraphs belonging to the sample discussed here.

[5] The corpus under the CLARIN.SI noSketch Engine is available at https://www.clarin.si/noske/run.cgi/corp_info?corpname=zrc16clt&struct_attr_stats=1. T

[6] The manual is available at https://www.sketchengine.eu/user-guide/.

degree of original variation (as described below). The majority of the remaining instances are also clitics, most frequently the negative proclitic *ne* 'not' followed by the enclitic particle *li* 'whether, if'. In some cases, enclitical conditional particle *bi* and reflexive particle *se* also form one orthographical word with their hosts, as do proclitical one-syllable prepositions *na* 'on', *ob* 'at, by', *pri* 'at, beside' and *za* 'for, behind'.

In individual instances two accented words are also written as one orthographical word. Most frequently those are phrases consisting of adjective + noun (e.g. *višji_far* '*high priest*', *figino_drevo* 'fig tree' and in reverse word order *gnade_bogat* 'merciful') or adverb + verb (*domov_iti* 'to go home', *zoper_stati* 'to resist'). As the analysed texts are translations from German where the quoted forms are compounds (*Hocherpriester* 'high priest', *Pfeigenbaum* 'fig tree', *gnadenreich* 'merciful'; *heimgehen* 'to go home', *wiederstehen* 'to resist') it can be assumed they were written together under the influence of German orthography.[7]

The absolute numbers of specific clitics partially correlate with the prevalence of bound variants in comparison with the freestanding variants of those clitics: the non-syllable prepositions predominantly form one orthographical word with their hosts (most notably *v* – in more than 90% of instances – while *k* is bound in approx. 70% and *z* in 70% of instances). Among other clitics the particle *li* is bound in 85% of instances and the negative *ne* is in approx. 40%, while the rest are freestanding in more than 90% of instances or their number is too small to be statistically relevant.

Word	TT 1557		JPo 1578		TPo 1595		All	
	bound	free	bound	free	bound	free	bound	free
v	49	0	105	0	78	19	238	19
k	23	0	73	0	35	53	127	53
z	12	1	61	0	47	77	120	78
li	1	1	46	2	0	5	47	8
ne	0	14	43	14	21	60	63	88
bi	0	15	9	34	0	82	9	131
se	2	20	2	73	0	120	4	213
na	0	23	0	62	3	67	3	152
ob	0	1	1	0	0	1	1	2
pri	0	2	1	7	0	9	1	18
za	0	13	4	15	2	39	6	67

Table 3. Most frequent bound words.

[7] For various mechanisms of borrowing from German into Slovene cf. eg. Legan Ravnikar 2017: 36.

Nearly all freestanding variants of non-syllable prepositions occur in TPo 1595. The only exception is the single instance of freestanding preposition *z* in TT 1557 but the variant used is one-syllable allomorph *zo* (see below).

However, even in TPo 1595 bound variants of *v* are highly predominant (approx. 80% of instances) while prepositions *z* in *k* are bound in approx. 40% of instances.

The majority of bound variants of other clitics can be found in other analysed work, JPo 1578. Only in that text are the enclitic *bi* and one-syllable prepositions *ob* (*obkratkim* 'in short') and *pri* (*priſebi* 'with himself') bound in some instances while predominantly remaining freestanding.

nepotrebuiesh/*ne_potrebuješ* **dabi/da_bi** te/*te* gdo/*kdo* kay/*kaj* vprashal/*vprašal*
NEG_need$_{2sg.}$ THAT_would you$_{ACCsg.}$ anyone anything asked
'You don't need that anyone asks you anything.'

Kadarbi/kadar_bi pak/*pa* tudi/*tudi* vſelei/*vselej* Salnce/*sonce* ſyalo/*sijalo* […]
when_would however also always the sun shone
'However, when the sun would always shine …'

In JPo 1578 we also find the majority of instances when negative particle *ne* is written together with its host word (in Juričič's text that spelling is three times more frequent than freestanding variant). Even more predominant is the bound variant of the particle *li* (which is freestanding only in 5% of instances) while the reflexive pronoun *se*, though in some instances bound, predominantly remains freestanding.

In TPo 1595 bound variants of one-syllable prepositions *na* 'on' (but limited to phrase *naznanje dati* 'to inform') and *za* 'for' (eg. *kaj zaen* /kaj za en/ as a calque translation for German *was für ein* 'what kind of') are used in individual instances but predominantly these prepositions are freestanding. The negative particle *ne* is bound in approx. 25% of cases.

In TT 1557 only individual instances of enclitics *li* and *se* are written together with their host words:

Hozhesli/*hočeš_li* popolnom/*popolnom* biti/*biti* […]
want$_{2sg._if}$ perfect to be
'If you want to be perfect …'

Vsdigniſſe/*vzdigni_se* , inu/*in* **uerfiſſe/vrzi_se** umurie/*v_morje*
lift_REXLEXIVE and throw_REXLEXIVE in_sea
'Lift yourself and throw yourself in the sea.'

3.2 One modernised word as two original words

While in cases where two modernised words form one original word clear patterns can be discovered, the reverse cases are sporadic or can be identified as errors in the original books. The one exception is the superlative adjective/adverb prefix *naj-/nar-* 'the most' that is orthographically bound with its root only in about 25% of instances. Of interest are also cases when beginnings of words that are homonymous with non- or one-syllable prepositions are separated from the remainder of the word with an apostrophe (eg. *s'_nameinja* 'signs', *s'_derſhati* 'to endure', *do_bruta* 'goodness', *sa_doſti* 'enough').

3.3 Variants of the preposition *k*

In the analysed corpus 5 spelling variants of the preposition *k* were attested. Besides bound variant <k_>[8] that appears in approx. 60% of instances spelling <h_> is used in around 10% of cases. /H/ as an allomorph of the preposition *k* is also used in modern day Slovene but it is limited to host words starting with a velar (/k/, /g/) while in the 16th century Slovene its use was positionally less limited as it could also be used in front of other plosives (in the analysed corpus it is attested in front of /b/, /p/ and /t/, but the variant <k> is used in front of those consonants with approximately equal frequency.[9] <h_> is the most frequent in TT 1557, while being rarely used in JPo 1578. In TPo 1595 the preposition with an apostrophe prevails for both variants, <k'> and <h'>. The latter is used not only in front of plosives /b/, /p/ and /t/ but also when the host word starts with affricates /c/ and /č/[10].

Another variant of the preposition *k* is <q_> that appears only in TT 1557 in front of the host starting with a *v-* (*quom* for *k vam* 'to you'; in one instance the usual spelling of the preposition <k_> can also be found – *kuam*). In general, in the analysed corpus the letter *q* appears only in front of *u* as a spelling for the consonant cluster /kv/ (eg. *cerque* 'churches').

3.4 Variants of the preposition *z*

The spelling of the preposition *z* shows similar variation. As *k*, it has two allomorphs: voiced /z/ used in front of voiced consonants and vowels and voiceless /s/ used in front of voiceless consonants. While the spelling of both /s/ and /z/ in the 16th century Slovene was highly inconsistent (<s>, <ſ>, <ſs> and <ſſ> were used for both while /s/ could also be written as <sſ> and <ß> and /z/ as <z> (Ahačič et al.

[8] Angle brackets indicate graphemes. Underscore indicates a bound variant.

[9] Merše, Jakopin and Novak (1992: 330) state that in Trubar's works <k> is predominant in front of vowels and sonorants while <h> prevales in front of voiceless plosives.

[10] In IPA these are t͡s and t͡ʃ; the first was generally graphically represented with <c> or <z>, the second with <zh> (Ahačič et al 2011: 29).

2011: 31); in the translation of the Bible <ſ> prevailed for /s/ and <s> for /z/) in the role of preposition and thus in word initial position only <s> and <ſ> were used[11].

In front of voiced consonants and vowels the preposition was usually spelled as <s>, <ſ> appears only once (*ſmyrom* 'in peace').

In cases where the host word starts with a voiceless consonant <ſ> is used more frequently, however <s> appears in more than a third of instances, most notably in JPo 1578 where it is consistently used in front of <k-> while <ſ> predominates in front of <t->[12].

Also in TPo 1595 where the preposition is predominantly separated from its host with an apostrophe (especially in front of voiceless consonants) only <s'> and not <ſ'> is used in such cases.

	TT 1557		JPo 1578		TPo 1595	
	voiced host	voiceless host	voiced host	voiceless host	voiced host	voiceless host
*ſ_*	0	7	0	13	1	3
*s_*	2	1	12	9	8	3
s'	0	0	0	0	34	42

Table 4. Spelling variants of the preposition *s*.

When the host begins with *s-/z-* the preposition, when written without the apostrophe, orthographically merged with the initial letter of the host word (eg. *sanashaniem* /z zanašanjem/, *ſramoto* 'with lenience', /s sramoto/ 'with shame').

Piſſari/*pisarji* ſo/*so* ſami/*sami* **ſebo**/**s_sabo** gouurili/*govorili*
scribes $AUX_{3pl.}$ with_REFLEXIVE PRONOUN spoke
'The scribes spoke amongst themselves.'

ie/*je* on/*on* [...] ta/*ta* Sueit/*svet* [...] **ſramoto**/**s_sramoto** [...] napolnil/*napolnil*
AUX3sg. he this world with_shame filled
'He filled this world with shame.'

[11] An analysis of the entire corpus confirmed that even outside prepositions, other variant spellings were rarely used in word initial position (eg. *ſsneg* 'snow', *ſſen* 'dream', *sſdaj* 'now'; <ß> was limited to a mid-word position and predominantly used between two vowels).

[12] Such distribution is also confirmed in the analysis of the entire corpus where in JPo 1578 <s-> in word initial position is only once used in front of <t> and <ſ-> never appears in front of <k>. In the other two analyzed texts such distribution is not attested; either grapheme is used in both positions.

kar/*kar* **sanashaniem/z_zanašanjem** samudi/*zamudi* [...]
what with_lenience misses$_{3sg.}$
'what he misses with lenience'

In front of palatal [ɲ] in oblique cases of the 3rd person personal and possessive pronoun the preposition *s* had a palatalized allomorph *ž* (IPA: ʒ; cf. Merše, Jakopin and Novak 1992: 331), which is non-existent in modern day standard Slovene. It was usually bound and spelled as <sh> or <ſh > (*shnymi*, *ſhnymi*). In TT 1557 and JPo 1578 <sh> is used while in TPo 1595 both spellings show similar frequency. In another spelling variant that occurs in TPo 1595 the digraph is parted by an apostrophe so that the <h> is orthographically attached to the host word: *s'_hyh* /ž njih/ 'with their'.

Only in one instance in TT 1557 variant <sh> is found in different context: *shkunshtio* /s kunštjo/ 'with skill':

[...] koku/*kako* bi/*bi* oni/*oni* Ieſuſa/*Jezusa* **shkunshtio/s_kunštjo** vieli/*ujeli*
how would they Jesus with_skill entrap
'[...] how they could skilfully entrap Jesus'

It is not clear whether the spelling <sh> in this case represents another allomorph /š/ (while not phonologically justified in front of /k/ it could be explained as the effect of trans-syllable assimilation) or it should be considered as another orthographical variant of the allomorph /s/.

The last two allomorphs of the preposition *s* attested in the analysed corpus are one-syllable /za/ and /zo/ that were limited to the cases where *s* preceded the adjective *ves* 'all, entire': /za/ appeared in JPo 1578 where it was bound and spelled as <ſa> (eg. *ſauſo* 'with all$_{ACCsg.\ f}$') while in TT 1557 and in TPo 1595 /zo/ was attested as a freestanding clitic and was spelled as <ſo> and <so>, respectively[13]:

opominai/*opominjaj* **ſauſo/z_vso** pohleunoſtio/*pohlevnostjo*
exhort IMPERATIVE2sg. with_all meekness
'exhort with all meekness'

Ieſt/*jaz* ſem/*sem* **ſo/z** uſo/*vso* dobro/*dobro* ueiſtio/*vestjo* hodil/*hodil*
I AUX1sg. with all good conscience walked
'I walked in good conscience'

ſluſhati/*slušati/,* inu/*in* **so/z** vſem/*vsem* fliſsom/*flisom* ſturiti/*storiti*
to listen and with all diligence to do
'to listen and to do it with due diligence'

[13] Cf. Merše, Jakopin and Novak (1992: 331). Variant /za/ is not mentioned as Juričič's work was not included in the analysis presented in the article.

3.5 Variants of the preposition *v*

The preposition *v* shows the highest degree of orthographical variation of all analysed words as it has – as a one-letter word – 10 different spellings (cf. Novak 2011). Partially, such inconsistency can be attributed to the long tradition of Latin scripts not distinguishing between the vowel /u/ and the consonant /v/ and using <u> and <v> for both sounds depending on their position (Pflughaupt 2008: 123–124). As <v> was traditionally used at the beginning of a word, the bound variant of the preposition <v_> was used in majority of cases in the analysed corpus (eg. *vdobri* 'in good', *vnebeſſa* 'to heaven', *vegypat* 'to Egypt'), followed by the bound variant <u_> (eg. *uiutro* 'in the morning', *umurie* 'to the sea', *unebeſih* 'in heaven') and the variant <v'_> separated from its host with an apostrophe (eg. *v'Pakal* 'to hell', *v'Paradiſhu* 'in paradise'). If the host word began with a vowel spelling variant <uv> (bound or separated with an apostrophe) was occasionally used (eg. *uven* 'in a $_{\text{ACCsg. m}}$', *uvozhito* 'in obvious', *uv'eni* 'in a $_{\text{LOCsg. f}}$'). In one case this variant crosses the word boundary as the apostrophe separates the <u> and the <v>, thus orthographically attaching the v to the host word: *u'_venim* = v_enem 'in a $_{\text{LOCsg. m}}$'.

Even more spelling variants can be found in cases when the host word starts with a v-. In such instances:

a) both the bound preposition and the host are spelled with <u> (<u_u...>, eg. *uuayu* 'in you two');
b) the bound preposition is spelled as <v> while the host is written with a <u> (<v_u>..., eg. *vuertu* 'in the garden');
c) the bound preposition is spelled as <u> while the host is written with a <v> (<u_v>..., npr. *uveri* 'in the faith');
č) the preposition is spelled as <u> and separated with an apostrophe from the host written with a <v> (<u'v…> *u'Voſi* 'in jail');
d) one-syllable variant of the preposition <va> is used with the host spelled with a <u> (<va_u...>, eg. *vauſeh* $_{\text{'in all LOCpl.}}$);
e) the preposition orthographically merges with the host into a single <v...> (<*vſej*> for v vsej $_{\text{'in all LOCsg. f'}}$);
f) the preposition orthographically merges with the host into a single <v> separated from the remainder of the host word with an apostrophe (<*v'ſem*> for v vsem 'in all $_{\text{LOCsg. m/n}}$).

	TT 1557	JPo 1578	TPo 1595
*v_*	23	85	52
*u_*	19	0	0
*v'_*	0	0	11
In front of a vowel			
*v_*	0	4	0
*v'_*	0	0	4
*uv_*	0	0	0
*uv'_*	0	0	1
u'_v	0	0	1
In front of v...			
u_u...	7	0	7
v_u...	0	6	2
u_v...	0	0	7
u'_v...	0	0	1
va_u...	0	10	0
v...	0	0	1
v'...	0	0	1

Table 5. Variants of the preposition *v*.

In TT 1557 bound variants <u_> and <v_> are used approximately equally, but when the host word starts with a *v*- in the analysed corpus only variant <u_> is attested (<u_u...>, eg. *uuas*).[14]

In JPo 1578 *v* shows even less variation, bound variant <v_> is used almost consistently, including in front of words starting with a v- (the initial v- of the host is consistently spelled as <u>: v_u..., eg. *vuoiski*); the only exception is when the host word is the adjective *ves* 'all, entire' in which case Juričič used the bound version <va_> (eg. *vaufei*).

In TPo 1595, on the contrary, almost all variants of the preposition described above appear: the most frequent being the bound <v_>, followed by non-bound <v'> with the apostrophe, bound variant <uv_> can be found in front of the hosts starting with a vowel and all quoted orthographical variants except a) and e) are used in front of a word starting with a v-. In that position the spelling <u_v...> is predominant, followed by <v_u...>, while other variants appear only in individual instances.[15]

[14] The result partially contradicts the findings of France Novak (2011: 129) who states that preposition *v* in that position is written as <v_u>. As the analysed corpus is limited to the hand-corrected sample determining the exact relationship of the spelling variants would require the normalisation and corpus analysis of the complete text of TT 1557.

[15] If we compare these findings to those in Novak 2011, we see that in our limited analyzed sample prevalence of bound variant <v> is confirmed, as well as spelling <u_v>

4 Conclusions

In the paper we have presented the construction and annotation of a corpus of samples from three Slovene books from the 16th century and, based on this corpus, performed an analysis concentrating on its clitics and clitic-like elements. We found that normalised corpus not only simplifies diachronic linguistic research, e.g. of word boundaries in the formative period of Slovene standard language compared to its modern norm, but also facilitates and enhances synchronic research of the 16th century Slovene literary language system. As the analysis of spelling variation in non-syllable prepositions showed, a relatively limited hand-corrected annotated sample enabled identification of majority[16] of spelling variants identified in previous works (cf. Merše, Jakopin and Novak 1992; Ahačič et al. 2011: 30–31; Novak 2011), while with the use of noSketch Engine tool further information about their relative frequency and distribution was obtained.[17] As the hand-corrected corpus is expanded such research will yield even more relevant information for the study of the 16th century Slovene literary language that will significantly supplement existing findings (based on traditionally collected examples) with the help of a large amount of statistically relevant data.

References

Ahačič, Kozma. 2014. *The History of Linguistic Thought and Language Use in 16th Century Slovenia*. Frankfurt am Main [etc.]: Peter Lang.

Ahačič, Kozma. Legan Ravnikar, Andreja. Merše, Majda. Narat, Jožica. Novak, France. 2011. *Besedje slovenskega knjižnega jezika 16. stoletja*. Ljubljana: Založba ZRC, ZRC SAZU.

---

and occasional <v_u> in front of *v*-, while Novak (2011: 138) does not mention frequent spelling <u_u> or orthographical merging of the preposition with the host word in that position in TPo 1595. On the other hand, in front of a vowel spelling <vu> (Novak 2011: 138) is not attested in the analyzed sample, as well as spelling <uv> that is illustrated in the article (the same spelling occurs with an apostrophe – <uv'>). Spelling <u'v> is identified in both.

[16] Excluded are spelling variants that (usually) do not occur in word-initial position, such as variants of *s* mentioned in 3.4 or <c> (mainly limited to words borrowed from Latin), <ck> (that only occasionaly occurs in word-initial position in TPo 1595 in proper name *Ckristus* for more predominant *Christus* 'Christ') and <g> for *k*.

[17] But it must be stressed that in corpus analysis one piece of typographical information that is in some cases relevant for the research of spelling is lost, i. e. the use of italic type that can in some cases account for the prevalence of one orthographical variant (cf. Novak 2006).

Eckart de Castilho, R. Biemann, C. Gurevych, I. Yimam, S.M. 2014. WebAnno: a flexible, web-based annotation tool for CLARIN. *Proceedings of the CLARIN Annual Conference (CAC) 2014*, Soesterberg, Netherlands.

Erjavec, Tomaž. 2015. *Reference corpus of historical Slovene goo300k 1.2*. Slovenian language resource repository CLARIN.SI, http://hdl.handle.net/11356/1025, 2015.

Erjavec, Tomaž. Jelovšek, Alenka. 2013. A corpus-based diachronic analysis of Slovene clitics. *New Methods in Historical Corpora*, 117–26. Tübingen: NarrVerlag.

Juričič, Jurij. 1578. *Poſtilla*. Ljubljana. Digital edition: *Korpus slovenskega knjižnega jezika 16. stoletja*, Jurij Juričič, *Poſtilla* [https://stage.termania.net/korpus16/].

Legan Ravnikar, Andreja. 2017. K problematiki vpliva stičnega jezika – nemščine na semantične spremembe in stilno vrednost najstarejše slovenske knjižne leksike (16. stoletje). *Slovenski jezik = Slovene Linguistic Studies* 11: 35–53.

Ljubešić, Nikola. Zupan, Katja. Fišer, Darja. Erjavec, Tomaž. 2016. Normalising Slovene data: historical texts vs. user-generated content. *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, 146–155.

Merše, Majda. Jakopin, Franc. Novak, France. 1992. Fonološki sistem knjižnega jezika slovenskih protestantov. *Slavistična revija* 40/4: 321–340.

Neweklowsky, Gerhard. 1985. *Das Werden* der slowenischen Schriftsprache. *Entstehung von Sprachen und Völkern*. *Glotto- und ethnogenetische Aspekte europäischer Sprachen*, 391–402. Tübingen: Max Niemeyer Verlag.

Novak, France. 2006. Predponi *v-* in *u-* v jeziku slovenskih protestantskih piscev 16. stoletja. *Stati inu obstati* 3–4: 138–159.

Novak, France. 2011. Predlog *v* v slovenskem knjižnem jeziku 16. stoletja. *Globinska moč besede: red. prof. dr. Martini Orožen ob 80-letnici*, 126–142. Maribor: Mednarodna založba Oddelka za slovanske jezike in književnosti, Filozofska fakulteta.

Pflughaupt, Laurent. 2008. *Letter by Letter: An Alphabetical Miscellany*. Trans. Gregory Bruhn. New York: Princeton Architectural Press.

Rigler, Jakob. 1968. *Začetki slovenskega knjižnega jezika*. Ljubljana: Slovenska akademija znanosti in umetnosti.

Scherrer, Yves. Ljubešić. Nikola. 2016. Automatic normalisation of the Swiss German ArchiMob corpus using character-level machine translation. *Proceedings of the 13th Conference on Natural Language Processing (KONVENS 2016)*, 248–255.

Toporišič, Jože. 2008/2009. »/S/ledni Sazhetig ie Teßhak inu nepopelnom«. *Slavistična revija* (*Trubarjeva številka*), letn. 56/57, št. 4, 1: 191–198.

Trubar, Primož. 1557. *Ta pervi deil tiga Noviga teſtamenta*. Tübingen. Digital edition: *Korpus slovenskega knjižnega jezika 16. stoletja*, Primož Trubar, *Ta pervi deil tiga Noviga teſtamenta* [https://stage.termania.net/korpus16/].

Trubar, Primož. 1595. *Hiſhna poſtilla*. Tübingen. Digital edition: *Korpus slovenskega knjižnega jezika 16. stoletja*, Primož Trubar, *Hiſhna poſtilla* [https://stage.termania.net/korpus16/].

Received April 2019, accepted June 2019.
Prispelo aprila 2019, sprejeto junija 2019.

Korpusna analiza klitik in njim podobnih elementov v slovenskem knjižnem jeziku 16. stoletja

V prispevku je predstavljena korpusna obravnava pisne variantnosti v slovenskem knjižnem jeziku 16. stoletja tako s sinhronega kot z diahronega vidika. Raziskava temelji na ročno pregledanem vzorcu (okoli 14.000 besednih enot) iz Trubarjevih del *Ta pervi deil tiga Noviga teſtamenta*, 1557, in *Hiſhna poſtilla*, 1595*,* in Juričičeve *Poſtille*, 1578, ter se osredotoča na klitike in njim podobne elemente. Statistična analiza zapisa skupaj in narazen na podlagi primerjave pisnih konvencij 16. stoletja in sodobne pisne rabe s pomočjo normaliziranih oblik izvirnih besed, tj. primerov, ko eni (ortografski) besedi iz 16. stoletja ustrezata dve ali več sodobnih besed (preslikava *1–n*) in obratno (preslikava *n–1*), je pokazala, da je tovrstnih primerov v korpusu 5,7 odstotka pojavitev, pri čemer so najpogostejše v JPo 1578 in najredkejše v TPo 1595, kjer jih je več kot enkrat manj kot pri Juričiču. V večini primerov gre za izvirno eno besedo, ki se v normalizirani obliki zapisuje kot dve besedi. Med besedami, ki so v izvirniku zapisane skupaj z gostiteljem, prevladujejo nezložni predlogi *v*, *k* in *z*, sledi jim členica *ne*, členek *li* in redkeje *bi*, *se* ter predlogi *na*, *ob*, *pri* in *za* (pri tem je njihova absolutna pogostnost večinoma premo sorazmerna z njihovo relativno pogostnostjo v primerjavi z njihovimi samostojno stoječimi variantami – najpogostejše so tudi prevladujoče zapisane skupaj z gostiteljem, pri redkejših pa prevladuje ločeni zapis). Posamezni primeri, ko sta združeno zapisani dve naglašeni besedi, so verjetno posledica vpliva nemških zloženk.

Zgledi, ko eni sodobni besedi ustrezata dve besedi v izvirniku, se pojavljajo sporadično – z izjemo presežniške prepone *naj-/nar-*, ki je v vzorcu zapisana združeno s pridevnikom/prislovom v okoli četrtini primerov – oziroma jih lahko označimo kot tiskovne napake ali napake v prepisu. Zanimivi so primeri, ko so vzglasni soglasniki homonimni z ne- ali enozložnimi predlogi in ločeni od preostanka besede z apostrofom (npr. *s'_nameinja*, *s'_derſhati*, *do_bruta*, *sa_doſti*). Normalizacija omogoča tudi ugotavljanje pisnih različic najpogostejših naslonk, ki se pisno povezujejo z naglašeno besedo, to je nezložnih predlogov *k*, *z* in *v*. *K* in njegov alomorf /h/ imata v korpusu izpričanih 5 pisnih variant, od katerih se <q_> pojavlja le pred besedami, ki se začnejo na *v-*. *Z* z zvenečim alomorfom /z/ in nezvenečim alomorfom /s/ ima v osnovi tri pisne različice, katerih razvrstitev se le delno ujema z (ne)zvenečnostjo prvega glasu sledeče besede, pojavljajo pa se tudi primeri pisnega zlivanja predloga s sledečo besedo, ki se začne s *s-/z-*. Druge pisne različice verjetno predstavljajo dodatne položajno vezane alomorfe: <sh/ſh/s'h> za mehčani /ž/ pred palatalnim *ń* in <ſa>, <ſo/so> za ozloženo različico /za/, /zo/. Predlog *v* izkazuje največjo mero pisne variantnosti med navedenimi

predlogi, saj ima kar 10 pisnih različic: splošni stično zapisani <v_> in <u_> in <v'_>, ki je v korpusu obravnavana kot nestična; <uv_>, <uv'_> in <u'_v> pred samoglasnikom; <u'_> in <va_>, ki se pojavljata samo pred *v-*; poleg tega pa še <v_> and <v'>, ki se pisno zlijeta z vzglasnim *v-* sledeče besede.

Analiza pisne variantnosti pri nezložnih predlogih je pokazala, da že relativno omejen ročno pregledan normaliziran korpus omogoča identificiranje večine pisnih različic, ki so navedene v obstoječi literaturi, uporaba programskega orodja noSketch Engine pa omogoča pridobivanje dodatnih podatkov o njihovi relativni pogostnosti in položajnih omejitvah. Z razširitvijo ročno pregledanega korpusa bodo tovrstne raziskave lahko prispevale pomembne in statistično relevantne podatke o sistemu slovenskega knjižnega jezika 16. stoletja, s katerimi se bodo dopolnjevale dosedanje ugotovitve, v glavnem temelječe na paberkovalnem izpisovanju primerov.

A corpus-based study of 16th-century Slovene clitics and clitic-like elements[18]

This paper undertakes a corpus-based linguistic investigation of the spelling variation in 16th century Slovene both from the diachronic and synchronic points of view. The investigation is based on a manually annotated sample (approx. 14,000 word tokens) from Primož Trubar's *Ta pervi deil tiga Noviga teſtamenta*, 1557, and *Hiſhna poſtilla*, 1595, and Jurij Juričič's *Poſtilla*, 1578, and it concentrates on clitics and clitic-like elements. Statistical analysis, based on comparison of the spelling conventions of the early modern period to those of contemporary Slovene using normalised forms of the originals, where we observe cases where one orthographic word is nowadays written as two or more words (1–n mapping) or vice-versa (n–1 mapping), shows that the overall percentage of split and joined word tokens is 5.7%, with JPo 1578 having the highest percentage, and TPo 1595 the lowest, less than half of that of JPo 1578. Of these, the vast majority is for cases where a word is now split. The most predominant among the bound words are non-syllable prepositions *v* 'in(to)', *k* 'to', and *z* 'with', followed by negative proclitic *ne* 'not', enclitic particle *li* 'whether, if' and in rare instances conditional particle *bi*, reflexive particle *se*, *na* 'on', *ob* 'at, by', *pri* 'at, beside' and *za* 'for, behind' (the absolute numbers of specific clitics partially correlate with the prevalence of bound variants in comparison with the freestanding variants of

[18] This article has been supported by ARRS (programs P6-0038 and P2-0103).

those clitics, with the most frequent being predominantly bound while the least frequent are predominantly freestanding). Individual instances of two accented words written together can be attributed to German influence (*figino_drevo, der Pfeigenbaum* 'fig tree').

The cases where one modernised word correlates to two original words are, with the exception of superlative adjective/adverb prefix *naj-/nar-* 'the most' that is orthographically bound with its root in about 25% of instances, sporadic or can be identified as errors in the original books. Of interest are also cases when beginnings of words that are homonymous with non- or one-syllable prepositions are separated from the remainder of the word with an apostrophe (eg. *s'_nameinja* 'signs', *s'_derſhati* '*t*o endure', *do_bruta* 'goodness', *sa_doſti* 'enough').

The normalisation also enables the identification of the orthographical variants of the most commonly bound clitics, i. e. non-syllable prepositions *k*, *z* and *v*. *K* and its allomorph /h/ have 5 attested spelling variants, of which one <q_> is limited to hosts starting with a *v*-. For *z* with a voiced allomorph /z/ and voiceless allomorph /s/ three variant spellings were discovered that only partially correspond with a voiceless/voiced distinction of the initial sound of the host word, and the cases of merging with the host that begins with *s-/z-* were identified. Additional positional spellings probably represent other allomorphs: <sh/ſh/s'h> for palatalized /ž/ in front of a palatal *ń* and <ſa>, >ſo/so> for syllabified /za/, /zo/. The preposition *v* shows the highest degree of orthographical variation of all analysed words as it has 10 different spellings: general bound <v_> and <u_> and freestanding <v'_>; <uv_>, <uv'_> and <u'_v> in front of a vowel; <u'_> and <va_> attested only in front of a *v*-, as well as <v_> and <v'> merged with the initial *v*- of the host.

The analysis of spelling variation in non-syllable prepositions showed that even a relatively limited hand-corrected annotated sample enabled identification of majority of spelling variants identified in previous works, while with the use of noSketch Engine tool further information about their relative frequency and distribution was obtained. As the hand-corrected corpus is expanded such research will yield even more relevant information for the study of the 16th century Slovene literary language that will significantly supplement existing findings (based on traditionally collected examples) with the help of a large amount of statistically relevant data.