Checking *POSBeseda*, a Part-of-Speech tagged Slovenian corpus

Birte Lönneker, Primož Jakopin

Corpus Laboratory, Fran Ramovš Institute for Slovenian Language
Scientific Research Center
Slovenian Academy of Sciences and Arts
Novi Trg 2, SI-1000 Ljubljana
birte.lonneker@guest.arnes.si, primoz.jakopin@uni-lj.si

Abstract

This paper presents recent observations concerning the Part-of-Speech (POS) part of *POSBeseda*, a POS-tagged Slovenian corpus at the Corpus Laboratory of the Institute for Slovenian Language ZRC SAZU. The tags include information on Parts of Speech (POS) as well as morphosyntactic descriptions (MSDs) and have their original foundations in Slovenian grammar. Special attention is paid to a precise description of the tagset, which allows for consistency checks of the tagging done so far. The checks are applied to *POSBeseda/2003* which contains slightly over 1 million words (excluding numbers and punctuation). An analysis shows that most inconsistencies in *POSBeseda* are due to different interpretations of the tagset, which result in under- or overspecified tags with regard to the tagset description presented in this paper. For those types of different interpretations that occur with a frequency of over 100, guidelines to resolve under- and overspecification are provided.

1. Introduction

POSBeseda is a Part-of-Speech (POS) tagged Slovenian corpus that has been developed at the Institute for Slovenian Language of ZRC SAZU in Ljubljana since 1995. The texts in *POSBeseda* were tagged by a tagger developed by the second author and verified and corrected by Aleksandra Bizjak (coauthor of the tagset and first Slovenian linguist to have dealt in depth with POS design issues for Slovenian) and Lučka Uršič, both from the Institute, using a POS-aware editor.[1] That verification procedure brought up many non-trivial questions which required a quick decision while the explanations in the reference books on Slovenian grammar were often inadequate or missing. The current version of the corpus (*POSBeseda/2003*) contains slightly over 1 million POS-tagged words. It is the only Slovenian POS-tagged corpus of this size and quality and will be used as a basis for automatic POS-tagging of the 121 million words *Nova beseda* corpus (http://bos.zrc-sazu.si), which is the main language resource produced and maintained by the Corpus Laboratory. The paper is in greater part based on observations gained by the first author when preparing resources for the application of an external tagger, the TreeTagger, to Slovenian texts.

Tags in *POSBeseda* include information on Parts of Speech (POS, such as noun, adjective, verb…) as well as morphosyntactical descriptions (MSDs) based on Slovenian grammar, which makes them very rich: Slovenian POS have many subtypes (especially pronouns), according to the reference grammars, and the MSDs provide a vast value list for inflectional features, e.g. six cases and three numbers (singular, dual, plural). Special attention is therefore paid to a precise description of the tagset, which allows for consistency checks of the tagging done so far. An analysis of *POSBeseda/2003* shows that there are some inconsistencies in the tagging, mainly due to different interpretations of the tagset, which result in under- or overspecified tags with regard to the tagset description presented in this paper. For those types of different interpretations that occur with a frequency of over 100, guidelines to resolve under- and overspecification are provided. The suggestions and opinions brought forward in the paper are aimed at easing the problem of large-scale POS tagging of Slovenian.

[1] An exhaustive tag dictionary did not exist at that time, as the tagset was expanded on the fly when relevant grammatical forms were encountered in text.

2. Motivation

The motivation for describing the tagset of *POSBeseda* and for checking the current version *POSBeseda/2003* is twofold. The first reason is that *POSBeseda* will be used as training corpus for tagging the entire *Nova beseda* corpus (121 million words). Some preliminary experiments were performed using the TreeTagger (Schmid, 1994), a statistical tagger. While the richness of the tagset is linguistically justified and very useful for linguistic analyses of the material, it was found to be an obstacle for the statistical tagging. An analysis of the tagging results showed as well that the tagging of *POSBeseda/2003* was inconsistent in some places; inconsistencies occurred especially in expressions for which also the different reference works on Slovenian grammar and lexicography propose a variety of interpretations.

A rule of thumb[2] for training the TreeTagger is that each tag of the tagset should have been used at least 100 times for the training to succeed; tags occurring between 10 and 99 times are of limited value; and tags with frequencies less than 10 are not suitable for the TreeTagger. Whereas statistical methods can anyway be successful only if the tagset does not exceed 150–200 tags, these frequency requirements were taken as an indication for splitting the tags occurring in *POSBeseda/2003* (including punctuation) into three frequency groups (see Table 1). It can be seen that the group of tags occurring more than 100 times is the

[2] Helmut Schmid, personal communication.

smallest group. We are thus thinking of methods to reduce the tagset, which would lead to a higher frequency of each tag.[3] The second reason for describing and analysing the *POSBeseda* tagset is to ease the manual correction of annotations produced by an existent POS-tagger for Slovenian (Jakopin, 2002). Ideally, the annotators have to be provided with detailed tagging guidelines similar to those of the STTS tagset for German (Schiller and Teufel, 1995). Such a set of guidelines would also lead to a higher consistency of the tagging. A by-product of this work is a list of all possible POS-tags including MSDs; it will be used in the future for simple checking procedures that would on the fly filter out spelling mistakes in tags introduced by the human annotators.

Occurrence limit	Distinct tags	Total occurrences
>= 100	432	1,230,430
< 100 && >= 10	750	26,319
< 10	883	2,887
Total:	2,065	1,259,636

Table 1: Occurrence classes in *POSBeseda/2003*

3. Description of the tagset

A thorough documentation of the tagset used has been given in Jakopin and Bizjak (1997). However, as an extension of this documentation has not been published yet, a consistency check has to involve a comparison of actual tag usage with the initial descriptions. To this aim, a list of all possible POS-tags including their MSDs has been created. The decisions underlying the creation process of this list are based on the presentation of the tags and their morphologic descriptions in Jakopin and Bizjak (1997); however, the precision of the model found in Jakopin and Bizjak (1997) has to be increased in some places.

3.1. Main inflection groups

According to the compositional pattern of the MSD, i.e. the available combinations of morphosyntactic description features, the following main inflection groups can be distinguished in POSBeseda/2003 (see Table 2):

I. Nouns (S), deverbal nouns (SG), proper nouns (IO–IL[4]), denominal adjectives (PIO–PIS), as well as most pronouns (ZV–ZSVP) and numerals (ŠV–ŠM) inflect for gender (3), number (3), and case (6). Also the reflexive possessive pronoun is included in this group, although the first case is rare to occur.[5] This MSD pattern pertains to 31 word classes (POS and subtypes of POS).

II. Verbs (G–GZ) inflect for person (3) and number (3). This MSD pattern pertains to eight word classes, including special word classes for present tense and future of the verb *biti* 'to be', and negated verbs (e.g. *ne imeti* → *nimam* 'I don't have').

III. Some participles (GN–PČ) inflect for gender (3), number (3), case (6), and (in the first and fourth case masculine singular) definiteness. This MSD pattern pertains to seven word classes.

IV. The past participle (GL, GLB) inflects for gender (3) and number (3). This MSD pattern pertains to two word classes, where GLB is reserved for the verb *biti*.

V. The imperative (GV) inflects for person (2; there is no imperative for the third person) and number (3).

VI. Adjectives (P) inflect for gender (3), number (3), case (6), degree (3), and definiteness (only marked in the first and fourth case masculine singular positive).

VII. Reflexive personal pronouns (ZOP) inflect for case (5; there is no reflexive personal pronoun for the nominative). The same pattern can be detected for prepositions (E), which are marked with the case they select (5; there is no preposition that selects the nominative). This MSD pattern pertains to two word classes.

VIII. Adverbs (A) can inflect for degree (3).

IX. Conjunctions (V) are subdivided into coordinating and subordinating conjunctions.

X. Both infinitive forms (GNE, GNA), the conjunctive (GBI), the participle on *-ši* (PŠI, GŠI), the free verbal morpheme (Gmp), the personal pronoun in conjunction usage (ZVR), the so-called indefinite numerals (ŠNE; e.g. *nekaj* 'some'), numbers (Š), the three kinds of particles (Č, ČZ, ČV), interjections (M), internet addresses (KURL), and abbreviations always followed by a full-stop, such as *ing.*, (KP) do not inflect. This MSD pattern thus pertains to 15 word classes.

XI. Some word classes require *special morphosyntactic description patterns*, which will be discussed separately in the next subsection (3.2.).

Table 2 indicates for each of these eleven groups based on MSD patterns the number of possible MSDs (column B), the number of word classes that inflect according to the described pattern (column C) and the total number of tags in the group (column D). It also lists for each group the word class component (POS component) of a POS-tag in *POSBeseda*, as well as an example and its translation into English.

3.2. Special inflection groups

The following word classes are not covered by any of the groups mentioned in Table 2: predicatives, personal pronouns, possessive pronouns, cardinal numbers, and abbreviations. The range of their tags is presented in what follows.

[3] A conceivable result might be to use two tagsets in parallel: A detailed one for manual tagging and a reduced one for statistical tagging.

[4] The word class IL is not present in the description given by Jakopin and Bizjak (1997); it pertains to Latin names like *Calamus aromaticus*.

[5] An example of a reflexive possessive pronoun in the first case is the following sentence from *POSBeseda*: *Saj mi je večkrat rekel, da bom na svoji zemlji* svoj *gospod!* 'As he told me several times, that I shall be my lord on my land.'

A	B	C	D	E	F	G
Description class	Number MSDs	Number word classes	Product B x C	Word class tags	Example	Translation
I	54	34	1836	S, SG,	dan (S)	day
				IO, IV, IB, IP, IN,	Martin (IO)	Martin
				IŽ, IZ, IM, IS, IL,		
				PIO, PIV, PIP,	Andrejev (PIO)	Andrej's
				PIZ, PIM, PIS,		
				ZV, ZR, ZPO, ZRPO,	kaj (ZV)	what
				ZNE, ZD, ZT, ZNI, ZI,		
				ZM, ZK, ZPU, ZSVP,		
				ŠV, ŠL, ŠM	prvi (ŠV)	first
II	9	8	72	G, GP, GFP, GZP,	plava (G)	swims
				GO, GZO, GFO, GZ		
III	56	7	392	GN, GT, GČ,	rojen (GN)	born
				PL, PN, PT, PČ		
IV	9	2	18	GL, GLB	pisal (GL)	(has) written
V	6	1	6	GV	glej	look!
VI	164	1	164	P	lep	beautiful
VII	5	2	10	ZOP, E	sebe (ZOP)	(oneself)
VIII	3	1	3	A	hitro	quickly
IX	2	1	2	V	in	and
X	1	15	15	GNE, GNA, GBI, PŠI,	povedati (GNE)	(to) tell
				GŠI, Gmp, ZVR, ŠNE, Š,		
				Č, ČZ, ČV, M, KP, KURL		

Table 2: Morphological description classes in *POSBeseda*.

Predicatives (PD). Most predicatives are invariable. However, following the Slovenian grammar (Toporišič, 2000, 412), also the word *rad* (used in expressions like *rad imam* 'I like') counts as predicative. It can only be used in the nominative case, where it inflects like an adjective without definiteness, so that descriptions of the following morphosyntactic features are possible: gender (3), number (3). The information on case is omitted in the tagset. *Rad* can also take the comparative and superlative degrees. While one form of them (*(naj)rajši, rajša, rajše*) is theoretically inflectable in the same way as the positive (cf. Toporišič et al. (2003)), there are parallel comparative and superlative forms (*(naj)raje*) which behave like adverbs and are never inflected for gender, number, or case. Accordingly, the following tags have to be allowed in the checklist: PD with gender (3), number (3) and degree (3); PD with degree (3), altogether **30** distinct tags.

Personal pronouns (ZO). All Slovenian personal pronouns inflect according to case (6). Information on person (3), number (3) and, in some cases, also on gender, are lexically integrated into the form. The third person singular is lexically marked for gender (3) in all cases, where the masculine and neuter gender often coincide (except for first case); the first and second person dual and plural distinguish masculine and feminine gender (2); and in the third person dual and plural in the nominative case, all three genders (3) are traditionally regarded as lexically distinct (Toporišič, 2000, 305–307), but those for feminine and neuter gender coincide in the dual number. The full range of tags with MSDs (**74**) for Slovenian personal pronouns is shown in Figure 1. Note that the tagset does not provide a distinction between clitic and non-clitic (accented) personal pronouns. Neither does it provide a special category for post-prepositional agglutinated pronouns (e.g. *vanj* 'into it'), which occur with prepositions selecting the fourth case: also those combinations have been tagged as pronouns.

Possessive pronouns (ZSV). All Slovenian personal pronouns, including the reflexive ones (ZSVP) mentioned in Subsection 3.1. above, inflect according to gender (3), number (3), and case (6) of the "possessed" noun. They could thus theoretically be regarded as belonging to inflection group I. However, the tagset also distinguishes between the person (3) and number (3) of the "possessing" entity, because this information is lexically present. In the tags, the lexical information precedes the inflection information (e.g. *moj*<*pos*>*ZSVaeme1*</*pos*>: *ae* pertains to the first person [a] singular [e] of the possessor, *me1* to the masculine gender [m], singular [e], nominative [1] of the possessed entity). The gender (2) of the possessing entity is marked only for the third person singular (e.g. *njegovi*<pos>ZSVcmemp1</pos>), where the gender of the possessor is male [m]). There are thus $3 * 3 * 6 * 3 * 3 + 3 * 3 * 6 =$ **540** MSDs for possessive pronouns.

Cardinal numerals (ŠG). Cardinal numerals can inflect for case (6) in Slovenian. The numeral *en* 'one' inflects for gender (3) and case (6); its number is semantically restricted to singular. The numeral *dva* 'two' is semantically restricted to dual number, while numerals higher than two are semantically of plural number. Inflections for gender

(3) occur only with the numerals *dva* in the first and fourth case and with *trije* 'three' and *štirje* 'four' in the nominative (feminine and neuter forms coincide). Starting with *pet* 'five', MSDs in *POSBeseda* have been reduced to case only (ŠG1, ŠG2, ...); however, cardinal numerals from five onwards can also occur without inflection and have no morphological descriptor (ŠG). The full range of POS-tags with MSDs (**43**) for cardinal numbers is shown in Figure 2.

Abbreviations (K). Although it is uncommon for abbreviations to inflect, some such cases have been found in the corpus, and annotated with a morphological description component. The occurrences were restricted to the singular number, but show inflection for gender (2; no neuter abbreviation was inflected) and case (6).[6] There are thus **13** morphological descriptors for K-abbreviations: K, Kme1, Kže1, Kme2, Kže2, etc.

ZOae1, ZOae2, ZOae3, ZOae4, ZOae5, ZOae6,
ZObe1, ZObe2, ZObe3, ZObe4, ZObe5, ZObe6,
ZOcme1, ZOcme2, ZOcme3, ZOcme4, ZOcme5, ZOcme6,
ZOcže1, ZOcže2, ZOcže3, ZOcže4, ZOcže5, ZOcže6,
ZOcse1, ZOcse2, ZOcse3, ZOcse4, ZOcse5, ZOcse6,
ZOamd1, ZOažd1, ZOad2, ZOad3, ZOad4, ZOad5, ZOad6,
ZObmd1, ZObžd1, ZObd2, ZObd3, ZObd4, ZObd5, ZObd6,
ZOcmd1, ZOčžd1, ZOcsd1, ZOcd2, ZOcd3, ZOcd4, ZOcd5, ZOcd6,
ZOamp1, ZOažp1, ZOap2, ZOap3, ZOap4, ZOap5, ZOap6,
ZObmp1, ZObžp1, ZObp2, ZObp3, ZObp4, ZObp5, ZObp6,
ZOcmp1, ZOčžp1, ZOcsp1, ZOcp2, ZOcp3, ZOcp4, ZOcp5, ZOcp6

Figure 1: The full range of morphologic descriptors for personal pronouns (ZO). 1st to 3rd person: a–c. Masculine m, feminine ž, neuter s. Singular e, dual d, plural p. Cases: 1–6.

ŠG, ŠG1, ŠG2, ŠG3, ŠG4, ŠG5, ŠG6,
ŠGme1, ŠGme2, ŠGme3, ŠGme4, ŠGme5, ŠGme6,
ŠGže1, ŠGže2, ŠGže3, ŠGže4, ŠGže5, ŠGže6,
ŠGse1, ŠGse2, ŠGse3, ŠGse4, ŠGse5, ŠGse6,
ŠGmd1, ŠGžd1, ŠGsd1, ŠGd2, ŠGd3,
ŠGmd4, ŠGžd4, ŠGsd4, ŠGd5, ŠGd6,
ŠGmp1, ŠGžp1, ŠGsp1, ŠGp2, ŠGp3, ŠGp4, ŠGp5, ŠGp6

Figure 2: The full range of morphologic descriptors for cardinal numbers (ŠG).

[6]The word class KI (abbreviation, always capitalized) mentioned in Jakopin and Bizjak (1997) is not in use any more.

3.3. Comparing the tagset with the corpus tags

From the presentation in the two previous subsections, it follows that the *POSBeseda* tagset is composed out of 3,218 tags including morphosyntactic descriptions. However, while not all of them actually occur in *POSBeseda/2003*, another 232 tags were found that are not described by this list, disregarding punctuation. There are several reasons for the additional tags. *Incorrect tags* may be caused by misspellings, when annotators were correcting the output of the previous statistical tagging procedure (cf. Section 1. above). *Underspecification* is achieved by leaving out one or more of the morphosyntactic description levels. It is a phenomenon introduced by human annotators in order to indicate dubious cases. An example is the expression *na*<*pos*>*E*</*pos*> *hitro* 'fast': here, the annotator could not decide which case is selected by the preposition *na* (it will be argued that the expression should be tagged as na<pos>E4</pos> hitro<pos>A</pos>, cf. Subsection 4.1.2.). *Overspecification* is in turn achieved by assigning a morphosyntactic description level to a form which does not contain this information.

Except in case of misspellings, these discrepancies show that the tagset was interpreted differently by the annotators – especially in the longer time frame – than in the interpretation presented above; therefore, individual contexts from the corpus will be discussed in order to illustrate the argumentation. Furthermore, word forms showing uncompliant tags somewhere in the corpus are very likely to be tagged with a compliant tag in the same or similar context somewhere else (for example, if *na* was once tagged as E and once as E4 in the same expression, *na hitro*). Therefore, all non-compliant tags are a possible source of artificial ambiguity, which would lower the precision of automatic tagging.

4. Checking *POSBeseda/2003* for consistency

The next step in checking the corpus consists in finding out the specific nature of the inconsistencies that were detected. To that purpose, the assigned uncompliant tags were divided into three frequency groups: those that occur 100 times or more – they are probably meaningful, but have been introduced because of a different interpretation of the tagset than the one presented here; those that occur between 10 and 99 times – they might be either meaningful, or due to a repeated error, possibly carried on by the automatic pretagging; and those that occur less than 10 times – they are very likely to be mistakes. For all tags, it has to be decided whether an adaptation of the model presented above is necessary, or otherwise in which way the uncompliant tags can be matched onto compliant versions.

In the remainder of this section, a method of analysis for uncompliant tags is presented. The explanations are based on 31 uncompliant tags that occur 100 times or more in *POSBeseda/2003*; they are due either to underspecification (11 tags) or overspecification (20 tags) with regard to the interpretation of the tagset that has been introduced above. The analysis of underspecified tags is presented in Subsection 4.1.; the treatment of overspecified tags is discussed in Subsection 4.2.

Occurrence limit	Uncompliant tags	Total occurrences
>= 100	31	25,935
< 100 && >= 10	55	2,110
< 10	146	373
Total	232	28,418

Table 3: Uncompliant tags in *POSBeseda/2003*

4.1. Underspecified tags

With four word classes, underspecification of tags occurs at least 100 times in *POSBeseda/2003*. These word classes are pronouns (15,528), prepositions (917), *č*-participle (206), and proper nouns for persons (121). While undespecification of the *č*-participle is a grammatically complicated matter that will have to be discussed elsewhere, the method of analysis for underspecified pronouns and prepositions is explained in detail in the following subsections (4.1.1.–4.1.2.); subsection 4.1.3. contains a brief remark on underspecified proper nouns.

4.1.1. Pronouns

With the most frequent underspecified pronouns, the entire morphosyntactic description is missing from the tag; for example, the tag ZT (so-called 'totality pronoun') occurs instead of ZTme1 (where the specifications for male gender, singular, nominative case are given). Pronoun tags in this group are ZČ, ZD, ZK, ZNE, ZNI, ZR, ZT, and ZV; each of them occurs more than 100 times in *POSBeseda/2003*. During the analysis of the corpus occurrences, a borderline wordclass was detected, about which there is some confusion in Slovenian grammar and lexicology and, not astonishingly, also in *POSBeseda/2003*. Most words that received the underspecified pronoun tag are indeed considered as adverbs in Bajec et al. (1970–1991), and as "adverbial pronouns" (*prisl. zaim.*) in Toporišič et al. (2003).[7] While semantic considerations might have led some linguists to consider these words as pronouns, this interpretation is not compliant with the given description of the tagset (see Section 3. above), in which pronouns inflect and thus have to be provided with an MSD component. Those underspecified words that do *not* inflect would have to be regarded as adverbs (occurring exclusively in the positive degree) or other non-inflecting words during tagging. Similar considerations are presented by Przepiórkowski and Woliński (2003), who devise a tagset for Polish based on morphological and morphosyntactic considerations only, leaving semantic interpretation (if needed) to a later processing stage. In what follows, analysis details for each of the above mentioned pronoun classes are presented.

ZČ. This word class (probably standing for 'time pronoun') is mentioned neither in Jakopin and Bizjak (1997) nor above. The only word bearing this tag is *tedaj* 'at that time' (247 occurrences). *Tedaj* has been interpreted as an adverb or "adverbial pronoun", but also as a particle or conjunction in Slovenian grammar and lexicology. It has however been tagged as ZČ throughout the corpus. It would be better to tag it as adverb (A) instead, because it does not show any inflection.

ZD. Following the definition given above, *drugje* 'somewhere else', *drugače* 'in a different way', and *drugam* 'somewhere else *(direction)*' are adverbs.

ZK. In this group, the following words do not inflect and will be regarded as adverbs: *tu*, *tule*, *tukaj*, *tod* 'here'; *sem*, *semle*, *semkaj* '(to) here'; *tam*, *tamle* 'there'; *tja*, *tjale*, *tjakaj* '(to) there'; *tak*, *tako*, *takó*, *takole*, *takóle* 'that way'; *toliko* 'so much, so many'; *odkod* 'from where'; *odtod* 'from here'. The re-interpretation as adverbs will be particularly useful here, because parallel annotations as adverbs already exist in *POSBeseda* for some of these words. The only word that might be a form of an inflecting pronoun is *tem<pos>ZK<pos>* (e.g. *s tem* 'with that'), a *demonstrative pronoun*. It occurs 17 times with an underspecified tag and would have to be checked manually.

ZNE. The words *malo*, *nekoliko* 'a little', *nekako* 'in some way', *nekam* 'to some place', *nekdaj* 'in old times; once', *nekje*, *nekod* (archaic) 'somewhere', and *veliko* 'much' have to be treated as adverbs. The word *kedaj* is an archaic form of *kdaj* 'when; sometimes', which has been tagged as ZV in *POSBeseda*, but should be regarded as an adverb (see below). However, *nekaj* 'some(thing)' is indeed a pronoun and inflects.[8] The underspecification (187 occurrences) is ambiguous between at least two[9] forms: nominative singular neuter (ZNEse1), found 278 times with the fully specified tag in *POSBeseda/2003* (e.g. *Nekaj<pos>ZNEse1</pos> na njegovem obrazu [...]* 'something on his face') and accusative singular neuter (ZNEse4), found 317 times (e.g. *[...] da nekaj<pos>ZNEse4</pos> zapiše* 'that he writes something down'). A disambiguation of *nekaj<pos>ZNE<pos>* (with underspecified tag) would again have to be performed manually.

ZNI. The following words have to be tagged as adverbs: *nikjer* 'nowhere', *nikdar*, *nikoli* 'never', *nikamor* '(to) nowhere', *nikakor*, *nikar* 'by no means'. The pronoun *nič*[10] 'nothing' (668 times underspecified) again shows an ambiguity between the first and fourth case neuter singular, which could be resolved only by manual checking.

ZR. The following words should not be tagged as pronouns: *kjer* 'there where', *kamor* 'where (to)', *koder* 'so far; (from) where', *dokler* 'as long as', *kadar* 'whenever, always when', *kolikor* 'as far as', *odkoder* 'from where', *od kar* (parallel form of the more frequent spelling *odkar*) 'since'. For *kadar* and *kolikor*, annotations as conjunctions (Vpo) also exist; the decision for the adverb tag might thus not be the optimal decision in this group, but further analysis would be necessary. *Kamor koli* and *kakor koli* have par-

[7] For illustration, the interested reader might look up the entries for *drugje*, *drugače*, and *drugam* in both reference works.

[8] *Nekaj* could be regarded as a member of a class corresponding to *indefi nite pronoun* in non-Slovenian grammars.

[9] According to the current tagset, it can also occur in the "indefinite numeral" reading (ŠNE); also this word class is based on semantic distinctions.

[10] *Nič* could also be regarded as indefinite pronoun.

allel (and more frequent) spellings *kamorkoli* ‘wherever’, *kakorkoli* ‘however’. These should also be tagged as adverbs. The pronoun *kar* ‘what; that which’ (similar to *relative pronouns* known also in grammars of other languages) shows an ambiguity between the first and fourth case neuter singular, which could be resolved manually.

ZT. In this category, all underspecified words are pronouns: *ves* ‘all; the whole’ (35 times underspecified) is ambiguous between first and fourth case masculine singular (ZTme1 and ZTme4). *Vse*, an inflected form of *ves*, is ambiguous between the following forms: ZTmp4, ZTse1, ZTse4, ZTže2, ZTžp1 and ZTžp4. *Vsake* (11), a form of *vsak* ‘every; each’, is ambiguous between ZTmp4, ZTžp1, ZTžp4 and ZTže2.

ZV. The following words are adverbs: *kam* ‘where to’, *kje*, *kod* ‘where’, *kako* ‘how’, *kolikokrat* ‘how many times’, *kdaj* ‘when’, *zakaj* ‘why’, *koliko* ‘how much; how many’, *doklej* ‘till when’. They are *interrogative adverbs*. However, *čemu* is a form of the pronoun *kdo* ‘who’ and ambiguous between the third and fifth case neuter singular (ZVse3 and ZVse5). Parallel annotations with full specifications exist. As the tag ZVse5 can occur only after an appropriate preposition (E/E5), rule-based disambiguation is feasible.

4.1.2. Prepositions

Prepositions (E) have sometimes been tagged without specification of the case they select, especially in contexts where they precede an uninflected form (or one that has been tagged as such). Here is an example: *na zadnji steni, od<pos>E</pos> koder<pos>ZR</pos> bi bil obvladoval vso sobo* ‘on the last wall, from where he would have had the whole room under control’. In this example, the underspecified preposition *od* is unambiguous with regard to the case it selects: It can only occur with the second case (cf. Table 4). It is straightforward to replace unambiguous prepositions with the respective morphosyntactically specified tag, which can be inferred from Table 4. These replacements are necessary especially because there are sometimes parallel annotations including morphosyntactic descriptions in *POSBeseda/2003*, as e.g. *od<pos>E2</pos> koder<pos>ZR</pos>*.[11]

For those underspecified prepositions that are ambiguous, some examples will be inspected more in detail. The most frequent among them are *na* ‘onto; on’, *po* ‘about; after’, and *za* ‘for; behind’.

Na. Some of the contexts in which the preposition *na* is underspecified can be classified as collocations in the sense of combinations of lexical items such that the semantics of one of them depends on the meaning of the entire collocation. For example, in *na hitro* ‘fast’, the preposition loses its central meaning ‘on’. This collocation type, in which *na* precedes an adverb, is quite common, cf. also *na pol* ‘halfway’; *na desno* ‘(to the) right’. The adverb can also occur in comparative degree: *[. . .] da gre svetu<pos>Sme3</pos> na<pos>E</pos> bolje<pos>Aj</pos>* ‘that the world is doing better’. These expressions are interpreted as showing a 4th case pattern by Toporišič et al. (2003, 893): “*na [. . .].* s tož. [. . .] razlagati na dolgo [. . .]; na tak način”. The preposition should thus be tagged as E4 in these collocations.

Other contexts are rare. An example worth mentioning is *na veliko načinov* ‘in many ways’: In analogy with *na tak način*, where the selection for the fourth case is marked, also this occurrence of *na* should be tagged as E4.

Preposition	Selected case(s)	Underspecifications
čez	4	3
za	4, 6	141
v	4, 5	15
s	6, 2	2
razen	2	2
proti	3	14
prek	2	1
pred	4, 6	1
po	4, 5	130
okrog	2	2
okoli	2	3
od	2	293
ob	4, 5	2
na	4, 5	214
nad	4, 6	1
med	4, 6	2
do	2	91
	Total	917

Table 4: Underspecified prepositions in *POSBeseda/2003*.

Po. The search for this underspecified preposition leads again to a number of interesting expressions. Two main groups of collocations can be distinguished: *po + adjective or possessive pronoun* and *po + numeral and other quantifying expression*.

In the first group (*po + adjective or possessive pronoun*), collocations of the type *po človeško* ‘in a human way’, *po hebrejsko* ‘in Hebrew’ are encountered, the second component of which is mostly interpreted as adverb in *POSBeseda/2003* (e.g. *po<pos>E</pos> hebrejsko<pos>A</pos>*). More consistent with Bajec et al. (1970–1991) and Toporišič et al. (2003, 1113) is the interpretation as a neuter adjective in the fourth case. The preposition can then be tagged as E4. In the correction procedure, attention has to be paid to the fact that not all adverbs following the underspecified preposition are in fact derived from an adjective; those would have to retain their adverb tag.

In the collocation type discussed so far, possessive pronouns can also be used, analogously to adjectives. The collocation type has not been interpreted unanimously in Slovenian linguistics, a fact which contributed to underspecifications in the corpus. For example, Bajec et al. (1970–1991) mark *po svoje* ‘one’s own way’ as an adverbial use of the reflexive possessive pronoun *svoj*, while Toporišič et al. (2003) interprets it as a collocation involving a form of the neuter noun *svoje*. *Svoje* in itself cannot be interpreted as an adverb, only the entire expression *po*

[11] Note also that the annotation of *koder* should be changed into koder<pos>A</pos> (see Subsection 4.1.1. above).

svoje could, but it would span token borders. For the tagging guidelines, a decision will thus have to be taken between tagging *svoje* as a reflexive possessive pronoun, or as a noun. Taking into account that forms of *svoje* have never been tagged as nouns in *POSBeseda*, it is proposed here to opt for *po<pos>E4</pos> svoje<pos>ZSVPse4</pos>*: The preposition selects the fourth case, and *svoje* is a reflexive possessive pronoun in the respective case. Similar collocations involve non-reflexive personal pronouns, of which two examples with the proposed tagging are given: *po<pos>E4</pos> moje<pos>ZSVaemp4</pos>* 'my way; according to my preferences'; *tisto zemljišče zdaj po<pos>E4</pos> njihovo<pos>ZSVcpse4</pos> imenuje Hakéldama* 'that plot of land he now calls, as they do it, Hakéldama'.

The second collocation group involving *po* could be called "collocations with numerals and other quantifying expressions". As can be seen from this description, a clear pattern that would cover all these collocations, based on morphosyntactic information only, is difficult to devise. What is more, it is not even possible to decide which case *po* should select in this group of collocations. In the following examples, the nouns and numerals following *po* have correctly been tagged as being in nominative case: *Zgoraj, na vsakem krogu, stoji po<pos>E</pos> ena<pos>ŠGže1</pos> Sirena<pos>Sže1</pos>* 'On top of each circle, (in groups of one) a Sirene stands'; *da se njeni člani lahko zberejo v večjem številu kot po<pos>E</pos> dva<pos>ŠGmd1</pos> ali trije <pos>ŠGmp1</pos>* 'that their members could gather in a higher number than (in groups of) two or three'. In other contexts, the noun or numeral is in the fourth case, which is clearly indicated by inflection in the following examples: *Navadni kmetje si lahko privoščijo po<pos>E</pos> eno<pos>ŠGže4</pos> ženo<pos>Sže4</pos>, bogati pa po<pos>E</pos> več<pos>Aj</pos>* 'Ordinary peasants can afford a single wife, and rich ones can afford more'; *držali so po<pos>E</pos> dve<pos>ŠGžd4</pos> ali tri<pos>ŠGžp4</pos> mere<pos>Sžp4</pos>* 'they held two or three measures'. Another point is that in most of these contexts, *po* could also be omitted, which is a strange behavior for a preposition. A solution might therefore be to regard *po* in these collocations as being an adverb (see Bajec et al. (1970 1991)), so that the selection preference could be omitted: *po<pos>A</pos> eno<pos>ŠGže4</pos> ženo<pos>Sže4</pos>*. However, this collocation type is one of the most difficult to deal with.

Za. Also the preposition *za* occurs in collocations describing quantifications, in combination with either an adverb or a noun phrase. If it occurs with a noun phrase, it selects the fourth case: *za<pos>E4</pos> dve<pos>ŠGsd4</pos> leti<pos>Ssd4</pos>* 'for two years'. The tagging guidelines should therefore indicate that also with adverbs, the selection preference of the prepositions should be marked as E4. Worth mentioning is also the expression *prav za prav* (a parallel spelling of the adverb *pravzaprav* 'actually'), which should be tagged as *prav<pos>A</pos> za<pos>E4</pos> prav<pos>A</pos>* instead of *prav<pos>Č</pos> za<pos>E</pos> prav<pos>Č</pos>*.

In other, rare contexts of underspecification, it is not possible to speak of a collocation involving the preposition. Such a case is *s<pos>E6</pos> svojim<pos>ZSVPse6</pos> za<pos>E</pos> in proti<pos>E</pos>* 'with his 'for' and 'against'', where both prepositions could be interpreted as the 6th case of morphologically invariable nouns (see Bajec et al. (1970–1991)); in order to keep the tagging simple, they can also be provided with the selection preferences they would show if followed by a noun in the respective context, i.e. *za<pos>E4</pos> in proti<pos>E3</pos>*.

4.1.3. Proper nouns for persons

In some cases, no morphosyntactic descriptions have been assigned to proper nouns for persons (IO), e.g. *Anne, Germaine*. This is the case only for foreign names. Some foreign names, especially female names not ending on *a*, are treated as invariant in Slovenian. Morphosyntactic specifications could therefore be assigned only with respect to the context in which the names occur, and underspecification could only be resolved manually.

4.2. Overspecified tags

Overspecification has been defined above as assigning a superfluous morphosyntactic description level to a form. In the frequency class of 100 and more occurrences, overspecification is encountered only with cardinal numbers and personal pronouns. For these word classes, special inflection patterns were detected and special MSDs were provided (see Section 3.2. above). The reduction of overspecifications is straightforward in most cases.

Cardinal numbers. Cardinal numbers are tagged with the word class tag ŠG. Two overspecified forms are found with a frequency of more than 100: ŠGmp4 and ŠGžp4. Here are two examples: *ura odbije tri<pos>ŠGžp4</pos>* 'the clock strikes three'; *kjer je delala za tri<pos>ŠGmp4</pos>* 'where she worked for three'. The proposed tagset implies that both tags should be replaced by the tag ŠGp4, in which no gender is specified.

Personal pronouns. Personal pronouns are tagged with the word class tag ZO; 18 overspecified forms occur with a frequency of over 100 in *POSBeseda/2003*: ZOamp3, ZOamp4, ZOamp5, ZOamp6, ZObmp2, ZObmp3, ZObmp4, ZObmp6, ZOcmd3, ZOcmd4, ZOcmp2, ZOcmp3, ZOcmp4, ZOcmp5, ZOcmp6, ZOcsp4, ZOčžp2, ZOčžp4. In total, there are 59 uncompliant tags for personal pronouns. Here are two examples: *[...] ki so jih<pos>ZOčžp4</pos> na Ministrstvu uporabljali* 'that they were using at the ministery'; *[...] ki jih<pos>ZOcmp4</pos> uporabljajo* 'that they are using'. The gender of the pronoun has to be inferred from its antecedent in the previous sentence. Semantically, the overspecified interpretation is thoroughly correct; however, there is no lexical or morphosyntactic way to detect it, so it is less suitable for POS-tagging (cf. also Przepiórkowski and Woliński (2003) for a similar argumentation). Some parallel forms – compliant with the tagset presented above – also accur, e.g. *jaz<pos>ZOae1</pos>* 'I', or *Vse okoli nas<pos>ZOap2</pos>* 'everything around

us'. The proposed tagset implies that all tags should be replaced by the corresponding tag without gender specification, i.e. ZOamp2 by ZOap2, ZOamp3 by ZOap3 etc.

5. Conclusion and future work

POS-tagged corpora for highly inflecting languages often show a multitude of tags that make them unsuitable for statistical training methods. The primary aim is thus to keep the tagset as small as possible. A first step towards this goal is certainly to keep the POS tags in the corpus "clean" (especially after the manual checking and correction phase following an initial automatical tagging), i.e. without uncompliant tags. The present article describes the application of this checking procedure on the Slovenian POS-tagged corpus *POSBeseda* and thus continues the recent line of work on the detection of errors in POS annotation (cf. Dickinson and Meurers (2003)). Our contribution shows that some morphosyntactically difficult phenomena are only encountered when the corpus reaches a certain size, and that solutions for these tagging problems can sometimes only be found in a POS-tagged corpus of considerable size, such as *POSBeseda/2003*, because a certain number of contexts is necessary to make the analysis possible. We analysed in detail most of the uncompliant tags of a frequency of 100 or more, and formulated tagging guidelines on the basis of a list of inflectional groups as well as on phenomena seen in the corpus.

As a remedy for many of the analysed inconsistencies, quite simple replacement procedures have been implemented using regular expressions. Preliminary experiments involve an application of these replacement procedures on *POSBeseda/2003* as well as on a test corpus of 200,000 words that was pre-tagged and hand-corrected in the same way as *POSBeseda*. The increased consistency of the corpora indeed had a positive influence on the precision of the results achieved using the TreeTagger. However, where rule-based replacement procedures for correcting the inconsistencies cannot be applied, they have to be corrected either manually, or using an improved statistical tagger based also on a future part of the corpus, which will be tagged according to the guidelines.

The checking and tagging of the full *Nova beseda* corpus will also be supported by a full-form dictionary of so-called closed class words, e.g. pronouns and conjunctions, which can be enumerated exhaustively (see also Dickinson and Meurers (2003) for the error detection method "closed class analysis"). Similar dictionaries have already been produced for Slovenian, for example in the frameworks of the Multext-East/Concede projects (Erjavec, 2004) and of the LC-STAR project (Verdonik et al., 2004).

Further reductions of the tagset might include the clustering of some of the word classes (i.e. the POS-components of the tags), especially of the various proper nouns, adjectives derived from proper nouns, and possibly pronouns. On the other hand, as far as the MSD component of the tags is concerned, results presented in Džeroski et al. (2000) show that the decision to abandon them or some of their features for training statistical taggers has to be well pondered. Using the Multext-East tagset, another full tagset for Slovenian, Džeroski et al. (2000) gained only 0.76% in absolute precision when tagging with the POS-component only, compared to tagging with the full tagset (including the MSD component) and predicting POS only.

The preparation of a machine POS tagger for Slovenian that would produce useful results on a sizeable amount of text is a complex task but, in the long term, also a very rewarding one. Many grammatical issues which in traditional linguistics received only sporadical attention, now come to the limelight. Their clarification would also be a major contribution to the completeness of the linguistic model of Slovenian.

6. References

A. Bajec, J. Jurančič, M. Klopčič, L. Legiša, S. Suhadolnik, and F. Tomšič, eds. 1970-1991. *Slovar slovenskega knjižnega jezika*. DZS.

M. Dickinson and W. D. Meurers. 2003. Detecting Errors in Part-of-Speech Annotation. In: *Proceedings of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL-03)*, pp. 107–114, Budapest, Hungary.

S. Džeroski, T. Erjavec, and J. Zavrel. 2000. Morphosyntactic Tagging of Slovene: Evaluating Taggers and Tagsets. In: *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC 2000)*, pp. 1099–1104, Athens, Greece.

T. Erjavec. 2004. MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In: *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC 2004), May 26–28, 2004*, pp. 1535–1538, Lisbon, Portugal. ELRA.

P. Jakopin and A. Bizjak. 1997. O strojno podprtem oblikoslovnem označevanju slovenskega besedila. *Slavistična revija*, 3–4:513–532.

P. Jakopin. 2002. *Entropija v slovenskih leposlovnih besedilih*. Založba ZRC.

A. Przepiórkowski and M. Woliński. 2003. The Unbearable Lightness of Tagging. A Case Study in Morphosyntactic Tagging of Polish. In: *Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora (LINC-03), EACL 2003*, Budapest, Hungary.

A. Schiller and S. Teufel. 1995. Guidelines für das Tagging deutscher Textcorpora mit STTS. Draft. Technical report, Universität Stuttgart and Universität Tübingen.

H. Schmid. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In: *Proceedings of the International Conference on New Methods in Language Processing*, Manchester, UK.

J. Toporišič, F. Jakopin, J. Moder, J. Dular, and S. Suhadolnik, eds. 2003. *Slovenski pravopis*. Založba ZRC.

J. Toporišič. 2000. *Slovenska slovnica*. Založba Obzorja Maribor, 4th ed.

D. Verdonik, M. Rojc, and Z. Kačič. 2004. Creating Slovenian Language Resources for Development of Speech-to-Speech Translation Components. In: *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004), May 26–28, 2004*, pp. 1399–1402, Lisbon, Portugal. ELRA.