Nina Ledinek
Šoštanj
Andreja Žele
The Fran Ramovš Institute of the Slovene Language in Ljubljana

Building of the Slovene Dependency Treebank Corpus According to the Prague Dependency Treebank Corpus

The article deals with the building of a syntactically annotated corpus of Slovene written texts – the Slovene Dependency Treebank – which is being modelled after the Prague Dependency Treebank. The proposed modifications of the surface-syntactic annotation system of the latter for the needs of Slovene will be illustrated by means of free verbal morphemes. We draw attention to, on the one hand, the high level of equivalence between Slovene and Czech syntactic phenomena and, on the other hand, the particularities of the Slovene language on the morphological and syntactic level.

Prispevek predstavlja projekt gradnje skladenjsko označenega korpusa slovenskih pisnih tekstov Slovene Dependency Treebank, ki nastaja po modelu korpusa Prague Dependency Treebank, in predlog za prilagoditev sistema površinskoskladenjskega označevanja zanj na primeru prostomorfemskih glagolov. Opozarjamo na visoko stopnjo prekrivnosti skladenjskih fenomenov slovenščine in češčine, hkrati pa izpostavljamo posebnosti slovenščine na morfološki in skladenjski ravni.

1 Slovene Dependency Treebank project

Syntactically annotated corpora are an important language resource, as they allow empirical syntactic analysis of language use patterns in large quantity of naturally occurring texts. Besides, they serve as a comprehensive internally unified and structured datasets, which can also be used for testing and training of automatic syntactic parsers. Data gathered on the basis of such annotated corpora is needed also for the development of a large number of language technology modules and tools. A syntactically annotated corpus of the Slovene language has not been available so far, although morpho-syntactically annotated and lemmatized corpora of Slovene are accessible, e.g. *FIDA* on <http://www.fida.net/slo/index.html> and *SVEZ-IJS* on <http://nl.ijs.si/svez/>).

However, the Slovene Dependency Treebank project (hereinafter: SDT) <http://nl.ijs.si/sdt/> was initiated at the Jožef Stefan Institute in 2003 and is taking place within the Department of Knowledge Technologies (Džeroski et al. 2006). The aim of this project is to build a syntactically annotated corpus of selected Slovene written texts. The theoretical basis of the annotation system is that of dependency syntax. At this point, the project is undergoing the phase of manually performed surface-syntactic annotation of a test corpus.

In terms of morphology the Slovene language is an inflectionally rich Slavic language with free word order (fixed word order mainly concerns the sequence of clitics in a string, a string of adjectives within a left attribute, functional words and rare subordinates, however, the word order itself depends mostly on the topicalization), hence the decision to take the Prague Dependency Treebank corpus (hereinafter: PDT) as a model for building the syntactically annotated corpus. This undertaking is one of the most ambitious and the best documented projects as far as syntactic annotation of languages, similar to Slovene, is concerned. Besides, a very large corpus covering three levels of annotation has already been made available for a comparative analysis. Due to the semantico-, functional- and structural-syntactic analogy of Czech and Slovene we were able to directly apply in our project not only a theoretical model, but also the surface-syntactic annotation system of the PDT corpus, defined in the manual *Annotations at Analytical Level: Instructions for Annotators* (Bémová, Alla et al.: 1999) (hereinafter: *AAL*).

The aim of the STD project is the development and analysis of an automatic syntactic annotation methodology of the Slovene language corpora, and the research of corpus data for the purpose of descriptive lingustics and language technologies development. The project is still at its early stage, and the researchers have two associated tasks to complete. The first one includes the comparison of surface-syntactic annotation manual with the existing descriptions of the Slovene syntax. The manual will then be adapted for the needs of the Slovene language, with regard to the experience and knowledge gained during manual annotation, and with regard to the understanding of the semantico-, functional- and structural-syntactic role ascribed to structures in contemporary Slovene linguistics. Later in the article it will be indicated how rules for the annotation of free verbal morphemes could be adjusted. The aim of the second task is the manual annotation of a test corpus. The latter will then serve as a dataset on a basis of which the adequacy of a surface-syntactic annotation system and software that were introduced will be estimated. The syntactic structure of each sentence is

represented by a syntactic tree structure in which the type of (a surface-syntactic) dependency of each token in relation to its direct governing node is defined. At this point the test corpus comprises approximately 1500 annotated sentences.

A Slovene part of a morpho-syntactically annotated parallel corpus MULTEXT-East <http://nl.ijs.si/ME/V3/> (Erjavec 2004) was chosen as a text for manual surface-syntactic annotation. The text is a translation of George Orwell's novel *1984*. The corpus is encoded in XML format and complies with the recommendations of Text Encoding Initiative TEI P4. It comprises approximately 100 000 tokens. This selection has some weaknesses (e.g. only one translated literary text serves as a basis for the corpus and even this one contains some invented language, the translation and the proof-reading of some parts of the text don't seem to be of the best quality) even so, the selection of a test corpus of this kind allowed the researchers to skip the morpho-syntactic phase of annotation. The latter was carried out with extreme precision as a disambiguation of morpho-syntactic functions and lemmas with regard to the context was made in two stages: the first one was carried out automatically with Eva text processor, after that the functions were hand-validated. Morpho-syntactic annotation system <http://nl.ijs.si/ME/V3/msd/> foresees approximately 2100 different annotations (for orientation: PDT corpus comprises 3000 different annotations). Consequently, when adjusting the *AAL* manual we have to consider a fact that a computer making a differentiation between syntactic structures uses somewhat smaller dataset concerning morphological categories of words in a sentence.

In later stages of the project we will focus, first and foremost, on three tasks. We will continue with the modification of the manual for surface-syntactic annotation (during the first stage we focused mainly on the adjustment of the annotation system for the structures, which in the Slovene linguistic are normally considered to be a predicate). The system will have to be changed in a way so that it will also define the annotation of the structures which are typical of Slovene. Besides, all Czech examples will have to be replaced with Slovene ones, while any other changes made in the *AAL* manual will have to be carefully documented. The scope of manually annotated sentences will be broadened as well. We will continue annotating the *»1984«* corpus and then focus on the compilation of a syntactically annotated corpus of approximately 200 000 words, consisting of different texts (especially newspaper articles and legal texts). The testing scope will be broadened together with the scope of improving the existent software applications for automatic syntactic annotation and of developing new ones.

In the year ahead we plan to update the project homepage. Additionaly, a small test corpus of manually annotated sentences will also be made available.

2 Surface-syntactic annotation of the SDT corpus from the linguistic point of view (limited to free verbal morphemes)

The comparison of contemporary descriptions of Slovene syntax with instructions for syntactic structure annotation of the Czech language in the *AAL* manual demonstrated that both languages have highly similar syntax, so most of the rules from the manual can be directly applied in the annotation procedure of the SDT corpus. Even so, from the point of view of the Slovene language we need to point to some of the manual's weaknesses which are mainly the result of the syntactic structure analysis going from the surface structure level towards meaning level, and of the simplifications, based on the automatic analysis of the language. This problem oriented approach presented next is confined mainly to a sample presentation of structures with verbs modified by free morphemes.

2.1 Free verbal morphemes

With the topic of free verbal morphemes we introduce the field of grammatical collocability as free verbal morphemes account for those morphemes that stand separately from the main part of a verbal lexeme and modify and determine different meanings of it. In the framework of Slovene linguistics these are divided into lexicalised free morphemes, forming a part of a lexeme meaning of a verb, and non-lexicalised free morphemes, merely accentuating the meaning of a verb on semantic and surface structure level. In Slovene three types of free verbal morphemes can be identified, namely pronominal, prepositional and personal pronominal morphemes.

The following part deals with a scheme for the adjustment of a surface-syntactic annotation system of free verbal morphemes in Slovene. From the point of view of the Slovene language the proposed solutions eliminate some of the weaknesses of the PDT corpus annotation system on both, surface- and semantico-syntactic level. Nevertheless, the annotation system preserves some of the surface-syntactic non-distinctive simplifications[1] as the automatic

[1] The term reffers to a different analytical functions used for the annotation of structures of the same type concerning the surface-syntactic level.

annotation of a surface-syntactic role of free morphemes proves to be very demanding. The fact is, lexicalised and non-lexicalised free morphemes cannot be distinguished on the basis of their surface structure and, furthermore, the distinction cannot be made on the basis of the (non)presence, surface structure properties and semantic features of (other) (non)participants of a predicate action and the structures denoting them.

2.1.1 Pronominal free verbal morphemes

All clitic forms of (initially) reflexive pronoun *se/si* (*self*$_{[accusative]}$/*self*$_{[dative]}$), regardless of their syntactic, semantic or morphological role, are regarded as pronominal free verbal morphemes. The definition is technical as the more detailed approach of these forms would prove to be too complex for the automatic language analysis. Furthermore, it is extremely difficult to define when a pronoun becomes a free morpheme, since this depends on semantic features of participants of a predicate action and on the context. The use of a term free morpheme for a clitic form of a pronoun of a type *umiti se, tepsti se* (*to wash oneself; to have a fight*) can also be justified by the fact that the pronoun only has a reference role in this case, i.e. a grammatical role of referring to a participant, usually assuming a role of the subject, and does not introduce any other participant.

2.1.1.1. The rules in the manual for surface-syntactic annotation are, in order to meet the criteria of automatic analysis of linguistic data, highly formalized and consequently take into account especially the surface structure level. However, in annotation of a surface-syntactic role of pronominal free verbal morphemes (the manual treats this type of free verbal morphemes exclusively) mainly the semantico-syntactic and semantic composition levels are taken into account. This orientation is demostrated by the use of a specific analytical function for lexicalised pronominal free verbal morphemes, by assigning different analytical functions to free verbal morphemes in surface-syntactic structures of the same kind in relation to the participants of the predicate action (*David se je premaknil – David moved; Veja se je premaknila – The branch moved*), etc. However, in the light of the development trend in reflexive pronouns and pronominal free verbal morphemes in Slovene this annotation system seems to be inadequate, since the tendency towards a more and more analytical manner of expression results in a free morpheme gradually adjusting to other elements of a semantic compositionality of a verb.

2.1.1.2 By taking into account, in particular, the assumed representation of free verbal morphemes on the semantico-syntactic level, the following analytical functions are proposed to be used for the annotation of a surface-syntactic role of pronominal free verbal morphemes in the Slovene language:[2]

1. Analytical function **AuxT**, as foreseen in the *AAL* manual, is being preserved for the annotation of lexicalised free morphemes *se/si*. It can be assigned only to a free morpheme that indubitably (in particular) assumes a lexical role (*smejati se – to laugh; delati se – to pretend; zdeti se – to seem; zapomniti si – to remember; domišljati si – to imagine,* etc.).
2. Analytical function **Obj** is proposed for the annotation of non-lexicalised rection-valent free morphemes of verbs, denoting reflexive actions. This analytical function can only be used in case when a free morpheme attached to a verb suggests that a subject actually acts upon itself (this is, as a rule, mostly the case with verbs denoting washing, dressing and taking care of one's appearance) (*umiti se – to wash oneself; obleči se – to dress oneself; tuširati se – to shower oneself,* etc.). Analytical function Obj is very seldom ascribed to a non-lexicalised rection-valent free morpheme *si* (*čestitati si – to congratulate oneself; škodovati si – to hurt oneself,* etc.).
3. Analytical function **AuxR** is proposed for the annotation of non-lexicalised free morphemes of verbs within typical sentence structures. This analytical function is assigned to free morphemes of verbs in passive structures (*Trava se kosi poleti – Grass is being cut in summer*), in structures with general doer of the action (*Govorilo se je o odkritju – A discovery was discussed*), in structures denoting uncontrolled (physiological) phenomena (*Zeha se mi – I feel the need to yawn*), in structures indicating a wish or a need to perform an action *(Pleše se mi – I feel like dancing)*, in typical expressions of colloquial use *(Išče se Uršo Plut – We are looking for Urša Plut*), in structures with »false doer of the action« (*Strižem se pri Miču – I have my hair cut at Mič's*), etc.
4. The technical analytical function **Atv** is proposed to be assigned to the rest of the non-lexicalised free morphemes *se* (*sprehajati se – to take a walk; skloniti se – to bend; spominjati se – to remember; postaviti se – to place oneself; utopiti se – to drown; ubiti se – to kill oneself; srečati se – to meet (each other); jeziti se – to be angry (with); Napetost se znižuje – The voltage drops; Veji sta se prepletli – The branches intertwined,* etc.).

[2] With the exception of technical analytical function Atv all other analytical functions are foreseen by the *AAL* manual. However, the ascription of surface-syntactic roles to various pronominal free verbal morphemes in the proposed annotation system differs from the system used for annotating Czech texts.

5. A non-lexicalised free morpheme *si* is assigned analytical function **Adv** except in cases when it is an obligatory rection-valent free verbal morpheme (*umiti si zobe – to wash one's teeth*; *Drevesi sta si stali nasproti – The trees stood opposite one another; izmenjati si čestitke – to exchange congratulations,* etc.).[3]

Since in the Slovene language both lexicalised and non-lexicalised free pronominal morphemes occur in participle as well as in gerundial (i.e. deverbative adverbial) structures, we suggest the same annotation system be used for free morphemes of verbal compounds.

The use of technical analytical function Atv for all non-lexicalised pronominal free verbal morphemes except for rection-valent morphemes of verbs, denoting reflexive actions and for free morphemes of verbs in typical sentence structures can be justified by the fact that these morphemes introduce pseudo-participants only. On account of different but in most cases high levels of semantic emptiness[4] of free morphemes, the syntactic structure of a verb and its free morpheme gradually morphologizes. This stage is actually a pre-lexicalization into other lexemes. Consequently, the morpheme *se* of a verb assumes a role of a grammatical/lexical morpheme and by that also the functional-syntactic role of a part of a predicate. However, *se* (*self*) preserves part of a referential meaning, namely »the reflexive one« (Žele 2003a: 17), thus implying the participant's existence, which can be felt in reduced valency of a verb. The boundary between lexicalised and non-lexicalised as well as between rection-valent (i.e. assuming the role of an actual participant of a predicate action) and non-rection-valent free verbal morphemes (i.e. having pseudo-participant role) is not clearly defined. In order to comply with a principle of consistency in annotation system, technical analytical function has been introduced for all border cases of free morphemes. By using this particular analytical function we can also avoid the possibility of assigning different functions to free morphemes of the same verb used in the same sense, merely on the basis of different conceptualizations of one and the same action (*Ubil se je v prometni nesreči – He was killed in an accident; Ubil ga je v prometni nesreči – He killed him in an accident*).

[3] When a morpheme is considered as redundant, the non-lexicalised pronominal free verbal morpheme can exceptionally be assigned analytical function AuxO. Structures of these kind are extremely rare in Slovene. Most often a morpheme of this type occurs in phrases that are border cases with fixed strings (*Bog si ga vedi (kdaj/kaj/kdo) – God knows when/what/who,* etc.).

[4] The term of semantic emptiness cannot be understood literally since it is difficult to establish the meaning of pronouns. A term »referential« emptiness would probably be more appropriate in this sense.

Analytical function Atv has been introduced also for non-lexicalised pronominal free morphemes of reciprocal verbs. This was also the case for morphemes (in contradiction to analytical function Obj foreseen by the *AAL* manual) of verbs, denoting reciprocal actions despite the fact that their free morphemes, when both participants are expressed with the subject, are rection-valent free morphemes. Structures with reciprocal verbs can denote various actions, even though they are homonymous on the surface structure level: they may account for parallel actions (*David in Hana sta se že poročila – David and Hana already got married = Both are married but not one with another*), collective actions (*David in Hana sta se srečala z učiteljem – David and Hana have met with their teacher = They had a meeting*), one-sided actions (*Vsak dan znova se srečujem s težavami – Every single day I am confronted with problems*), reciprocal actions (*David in Hana se tepeta – David and Hana are having a fight = They beat one another*), situational relations between the objects (*Veji sta se zapletli – The branches intertwined*) and various combinations of just mentioned actions (*Zdravnika se srečujeta s strokovnimi problemi – The doctors are facing technical problems*) (Shigemori Bučar 1992b). In these structures a free morpheme does not always occupy the position that would normally be occupied by the (co)actor of the action. Instead, the morpheme assumes a role of to a certain extent semantically emptied grammatical marker, designating predicate actions of various kind, hence its different level of participancy.

We opted for a surface-syntactic non-distinctive simplification in annotation procedure as, for the time being, automatic analysis of the language does not allow to differentiate between pronominal free morphemes of reciprocal verbs with different functions. These can frequently be distinguished solely on a basis of the context. In relation to the semantic features of participants of a predicate action (animate+/-, human+/-, abstract/concrete) we could distinguish, by means of a valency dictionary, only between free morphemes of those reciprocal verbs, denoting one-sided actions (*Vsak dan se srečujem s težavami – I face problems on every day basis; Grdo se gledam z računalnikom – I am not on good terms with my computer*) and those, being a combination of one-sided actions and actions of other types (*David in Lija sta se mučila z avtomobilom, ker ni hotel vžgati – David and Lija had problems with the car because it would not start*), as only in these structures with reciprocal verbs one of the participants occupying a place that would otherwise be occupied by the (co)agent of the »reciprocal action«, is inanimate. Since this participant, which is always denoted by an object in instrumental case cannot be a (co)agent of a predicate action the free morpheme of reciprocal verb predominately assumes a grammatical role. However, since the instrumental

case from the point of view of sentence elements makes it extremely difficult to draw a line between objects and adverbial adjuncts (Žele 2001: 96), the actions, similar to one-sided actions (*Spoznala sta se z internetom – They became acquainted with the internet; Tepel sem se z nožem – I fought with a knife = using a knife; David in Lija se srečujeta z veseljem – David and Lija like meeting each other; Boril sem se s samim seboj – I fought with myself = used metaphorically*) cannot be identified during automatic analysis of the language. Consequently, we suggest that free morphemes of reciprocal verbs, designating one-sided actions, also get the technical analytical function Atv.

Analytical function Atv for free morphemes of reciprocal verbs assuming a role of actual participants of reciprocal actions, was chosen with the purpose of annotation system simplification. *AAL* manual foresees analytical function Obj for these free morphemes, which means, the pronominal free verbal morpheme always assumes a role of an obligatory valency complement. However, the annotation of this kind may be problematic when annotating structures in which one of the participants on a surface-syntactic level occupies the place of the subject and the other that of the object in instrumental case. Therefore, despite the fact that there are only two participants of a predicate action (*David se je srečal z Lijo – David met with Lija)*, three obligatory valency places and thus the same number of obligatory complements (i.e. two object complements and one subject complement) can be identified on a surface-syntactic level. Since the structure in which both of the main participants of a reciprocal action are represented by a subject, is actually a transformed version of the above mentioned structure, we suggest that a unified annotation system be used. This means that all pronominal free morphemes of reciprocal verbs are assigned analytical function Atv. This annotation system can be justified also by the fact that lexicalised free morphemes of reciprocal verbs when assuming a role of actual participants of predicate actions are given analytical function AuxT (*Otroka se pogosto prepirata – The children often argue (with one another); David in Lija se borita za prevlado – David and Lija fight for the supremacy = They fight one another*).

The proposed annotation system allows that on both, surface- and semantico-syntactic level actual participants of predicate actions denoted as objects and pseudo-participants, which on a surface-syntactic level form a part of a predicate, can be distinguished on a basis of a small degree surface-syntactic (and semantico-syntactic) non-distinctive simplification. However, at a later stage of the annotation procedure a semantico-syntactic function (i.e. tectogrammatical

label) of rection-valent free morphemes of reciprocal verbs denoting actual reciprocal actions could be corrected manually on a small sample corpus.

We introduce a specific analytical function AuxR for the annotation of non-lexicalised pronominal free morphemes of verbs in typical sentence structures, in order to draw attention to the specific role of this morpheme. This is hardly an example of a typical morphological or lexical free morpheme of a verb (although a free morpheme of verbs in, particularly, passive structures, to some extent preserves the reflexive pronoun's initial function of introducing a new participant of a predicate action). Instead, it is considered as a syntactic grammatical morpheme or a modal label, which functioning merely on a surface structure level modifies the above mentioned sentence structures by attributing them explicit modal properties. In case we want them to preserve other meanings apart from a denotative one, actions designated by these structures can only be represented with one surface structure pattern, as a pronominal free morpheme drives any action of this kind in the direction of eventness.

Taking into account the valency properties of verbs and the specificity of syntactic structures, free morphemes with analytical function AuxR can, for the most part, be distinguished automatically. The fact that the range of verbs with free morphemes that can function in *se* sentence structures is limited also facilitates the annotation procedure.

2.1.2 Prepositional free verbal morphemes

Syntactic annotation of prepositional free verbal morphemes is with regard to the functional-syntactic level quite demanding, since they introduce objects and adverbial adjuncts, respectively. Lexicalised ones form part of a predicate on a functional-syntactic level, while non-lexicalised ones form part of a valent object or adverbial adjunct.

For the annotation of prepositions and prepositional free verbal morphemes the *AAL* manual anticipates one analytical function only, i.e. AuxP. Respecting the principle of consistence in the annotation procedure on a surface-syntactic level, it is therefore impossible to distinguish between different lexemes. Disambiguation would only be possible in case of consistent differentiation between prepositions as lexicalised (*hoditi z/s – Že tri leta hodi z njo – He goes out with her for the past three years; imeti za – Ima jo za bogato – He considers her rich*) and non-lexicalised rection-valent free verbal morphemes, respectively (*hoditi + na/v/skozi/čez/po*

čem and similarly – *Hodi na tečaj francoščine – He attends a French course*) and non-valent prepositions (*hoditi* (*od–do*, *po*, *med* and similarly) – *Hodil je po sobi – He walked around the room*, *Hodi po prstih – He tiptoes*).

Given the manual's surface-syntactic annotation system, this distinction being inconsistent, we suggest a technical solution, i.e. the assigning of a specific analytical function only to lexicalised prepositional free verbal morphemes.[5] By this we introduce the uniformity of the annotation system, since this latter already foresees a specific analytical function for lexicalised pronominal free verbal morpheme, as opposed to analytical functions for non-lexicalised one. Despite the fact that such an annotation system represents a non-distinctive simplification on a surface- and semantico-syntactic level, we suggest a similar system be used also for the annotation of prepositional free verbal morphemes. While lexicalised ones would be assigned analytical function AuxT (*stati za – David stoji za svojo odločitvijo – David stands behind his decision*), the non-lexicalised ones would preserve analytical function AuxP *(stanovati – David stanuje v Šiški – He lives in Šiška*). In this way, the surface-syntactic annotation of predicates and structures, following prepositions and free verbal morphemes, respectively, will be more accurate, since the lexicalised morphemes can only be followed by objects.

The automatic disambiguation between lexicalised and non-lexicalised prepositional free verbal morphemes and prepositions is no more demanding than the annotation of pronominal free verbal morphemes. However, it will only be feasible when a computer has access to a valency dictionary, as in many cases the disambiguation cannot be predicted on a basis of a surface structure level. However, the valency dictionary of Slovene does not exist for the time being, the first step in this direction represents a valency manual which by means of sample entries points to the typology of verbal valency in Slovene (Žele 2003a). With regard to the fact that a disambiguation of structures with verbs modified by prepositional free morphemes can often be made solely on the basis of the context (*prepirati se za hišo* 'prepirati se v zvezi s hišo' – *to argue over the house* – ali 'prepirati se zadaj, za hišo' – *to argue behind the house*), and that it cannot be made on the basis of data on surface structure level and on structure-syntactic valency, which a computer has at its disposal, we assume, the annotation of

[5] Since non-lexicalised rection-valent prepositional free verbal morphemes and non-valent prepositions never assume their own functional-syntactic role, the use of the same analytical function for both does not present a problem (as opposed to non-distinctive annotation of pronominal free verbal morphemes).

prepositional free verbal morphemes would in large part have to be performed manually. The proposed annotation of prepositional free verbal morphemes is thus planned for later stages of syntactic annotation and will probably be performed on a relatively small sample corpus.

2.1.3 Personal pronominal free verbal morphemes

Personal pronominal free verbal morphemes occur in the Slovene language in a specific type of verbal phrasemes or idiomatic verbs[6] (due to the specificity of the structure, the status of free morphemes (i.e. *jo, ga, jih – her, him, them*) cannot be formally proven for the majority of phraseological units, it is introduced according to the analogy with other free verbal morphemes – these clitic personal pronouns will be regarded as free morphemes in this section as well) with the »internal« verb + clitic form of the personal pronoun structure (*pobrisati jo – to make off; zadeti se ga – to get high; žurati ga – to party,* etc.). Their free morphemes have the role of non-rection-valent accusative complements, while on the functional-syntactic level they form a part of a composed predicate.

The basic differentiation between lexicalised and non-lexicalised free verbal morphemes, which was proposed for the annotation of other free verbal morphemes cannot be introduced for the annotation of personal pronominal free morphemes, since they are all lexicalised, however, it is impossible to establish whether their lexicalisation took place as a part of multi-word phraseological units or as a part of idiomatic verbs. The verbs in the above mentioned phraseological units (according to their »internal« structure) are mostly characterized by the so called absolute valency[7] (i.e. absolute semantico-syntactic use), since in all of their senses they are predominately rightward-valent. Thus, with regard to the semantic and surface structure properties of the participants of predicate actions denoted by the respective verbs, it is often impossible, by means of an automatic analysis, to distinguish between homonymous structures of verbal phrasemes (*pobrisati jo* 'uiti'– *to make off*) and phrases consisting of a verb and a personal pronoun (*pobrisati jo* 'pobrisati jo (tablo)' – *to clean it (the board) = to clean the board)*). Automatic analysis would allow the identification of only a small portion

[6] From the point of view of syntactic annotation of the corpus, the most important question, in relation to fixed expressions such as verb + personal pronoun in its clitic form, is the following: are they multi-word phrasemes or idiomatic words? Slovene linguistics does not provide a final answer to this question, due to the abundance of arguments, supporting each one of the respective theories. Generally, the level of idiomaticity or motivation and (non)inflectability of personal pronouns when used in negative sentences, are used as a standard for distinction between words and phrases.

[7] In a valency sense, these phraseological units act as phrases.

of free morphemes of the verbs of which the model monocollocability is typical – i.e. free morphemes of verbs which are rightward-valent only in a phraseological unit, and free morphemes of those verbs, where the collocation with the personal pronominal free morpheme constitutes the only possible and obligatory choice.

This is why we propose analytical function Obj be assigned to personal pronominal free verbal morphemes of verbal phrasemes. The annotation system of this kind is again a surface-syntactic non-distinctive simplification. However, in this way, the annotation system will be more consistent, since the elements of semantically and functionally identical structures will be assigned identical analytical functions. Apart from that, this kind of annotation system is also justified by the fact that it is virtually impossible to state with certainty that verbal phrasemes are in fact words. Actually, corpus data shows (ex. Kržišnik 2004: <http://www-gewi.kfunigraz.ac.at/gralis/GraLiS%202004/Krzisnik%20Frazemy.htm>) that personal pronominal elements of phraseological units (at least in written texts) are still subject to the rules of syntax, since the case of personal pronominal free morpheme is changed when used in negative sentences.

The proposed annotation is problematic especially from the point of view of semantico-syntactic annotation, but we presume the level of non-distinctive simplification will not be substantial, since the verbal phrasemes are generally rare in written texts. Therefore, their frequency in corpora is also expected to be low.

2.2 Slovene Dependency Treebank corpus annotation and Slovene linguistics

A contrastive analysis of descriptions of syntactic structures in the *AAL* manual and contemporary Slovene linguistics has proven to be very useful, since a mutual improvement of linguistic descriptions is possible due to the similarities between Slovene and Czech. The *AAL* manual, which contains a relatively comprehensive overview of syntactic structures in Czech, is interesting from linguistic point of view as well, since via its modification we will be able to obtain data about the extent of the discrepancies between the Slovene and the Czech linguistic system, as well as data showing to which degree these differences in description result from a different interpretation of the same structures. The comparison has demonstrated only a partial accordance of both languages on the level of morphemes, whereas a degree of the similarity on the syntactic level is significantly higher. The majority of

differences in descriptions are the result of taking into account the established models of linguistic theory. The most interesting discrepancies in descriptions are seen in the classification of non-valent complements into different functional-syntactic roles. We point to the differences in defining the role of non-valent free datives (in Slovene these are sometimes assigned a role of the object, as opposed to the description in the *AAL* manual, whereas the division of datives to non-valent and (non)obligatory valent complements varies according to different authors), gerunds (i.e. deverbative adverbs) (in Slovene they are considered as adverbial adjuncts and in Czech as complements, i.e. verbal attributes), »compound verb forms« (i.e. verbal phrases consisting of verbs only) (Slovene linguistic descriptions sometimes define them as being composed of copula and subject complement while in the *AAL* manual they are regarded as verb + object structures), etc. The typology of the adverbial adjuncts and inclusion of structures among them varies as well. Simplifications and the use of technical analytical functions are not taken into account (e.g. qualification merely of the verb *biti* as a copula, assigning of adverbial function to prepositional phrases, interjections and adverbs accompanying the verb *biti*) since they are imperative, given the automatic analysis of the syntactic structures. However, these simplifications and functions do not always correspond to the traditional linguistic descriptions on account of their non-computational orientation.

3 Conclusion

Surface-syntactic annotation of free verbal morphemes represents one of the more demanding and time-consuming tasks in automatic analysis of the language. The semantico-grammatical role of prepositions, pronouns or free morphemes accompanying a verb can be predicted mainly on the basis of semantic features of an individual verb and from the context, as well as on the basis of semantic features of participants of a predicate action. Therefore, the forseen annotation system (with regard to the structure type in which verbs with free morphemes occur, actions denoted by these verbs, etc.) is not highly appropriate if we want to define with precision the syntactic role free morphemes have in different structures. Even so, considering the extremely demanding task of building a syntactically annotated corpus, this seems to be the only feasible solution. This is why, at this stage, the rules for the annotation of free verbal morphemes represent a set of compromises, and the degree of surface-syntactic non-distinctive simplification is relatively high. The problem of the annotation will have to be tackled gradually, at several stages, while semantic data in the form of a dictionary will have

to be available to the computer. Nonetheless, by analysing the Slovene Dependency Treebank corpus, we expect to get highly reliable data on the occurence patterns of certain syntactic structure sets that have not been available to the Slovene linguistics so far. With the syntactically annotated corpus of Slovene at its development stage, we can help to foster the development of Slovene linguistics, at the very least by pointing to numerous structures that have not been subject to a lingustic description so far.

References

Bémová, A. et al. (1999): *Annotations at Analytical Level: Instructions for Annotators*. Praga: UK MFF UFAL.
<http://quest.ms.mff.cuni.cz/pdt/Corpora/PDT_1.0/References/aman_en.pdf>.

Džeroski, S. et al. (2006): Towards a Slovene Dependency Treebank. In: *Proceedings of the Fifth International Conference on Language Resources and Evaluation, LREC'06*. [Submitted].

Erjavec, T. (2004): MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In: *Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC'04*. ELRA: Paris, 1535–1538. <http://nl.ijs.si/ME/V3/>.

Golden, M. (1996): *O jeziku in jezikoslovju [About Language and Linguistics]*. Ljubljana: Faculty of Arts, Department of Comparative and General Linguistics.

Korošec, T. (1977): Slovenski dajalnik in dajalniške pretvorbe [The dative in Slovene and Its Transformations). In: *13. Seminar slovenskega jezika, literature in kulture [The 13th Seminar on the Slovene Language, Literature and Culture]*. Ljubljana: Department of Slavic Languages and Literature at the Faculty of Arts in Ljubljana, 59–67.

Kržišnik, E. (2001): Frazemi s strukturo »glagol + osebni zaimek« v slovenskem jeziku [Phrasemes with »Verb + Personal Pronoun« Structure in the Slovene Language]. In: *Frazeografija słowiańska*. Opole: Uniwersytet Opolski, 239–248.

Ledinek, N. (2005): *Površinskoskladenjsko označevanje korpusa Slovene Dependency Treebank (s poudarkom na predikatu) [(Surface-Syntactic Annotation of the Slovene Dependency Treebank Corpus (with Focus on the Predicate)]*. B. A. thesis. Ljubljana: Faculty of Arts, University of Ljubljana.

Orešnik, J. (1996): Nauk novejše slovenistike o povedkovem prilastku [Depictive Secondary Predication in Recent Slovene Linguistics]. *Razprave SAZU XV [Dissertations SAZU XV]*. Ljubljana: Slovene Academy of Sciences and Arts, 255–267.

Pogorelec, B. (1968): Razvoj prostega stavka v slovenskem knjižnem jeziku (Vloga dativa v stavku) [Development of a Free Sentence in Slovene Literary Language (The function of dative in a sentence)]. *Jezik in slovstvo* 13, 145–150.

Shigemori Bučar, C. (1992a): Izražanje povratnega dejanja v japonščini in slovenščini [The Expression of Reflexive Action in Japanese and Slovene]. *Slavistična revija* 40, 143–157.

Shigemori Bučar, C. (1992b): Izražanje vzajemnega dejanja v slovenščini in japonščini [The Expression of Reciprocal Action in Slovene and Japanese]. *Slavistična revija* 40, 365–383.

Shigemori Bučar, C. (1993): Izražanje samodejnega dejanja v japonščini in slovenščini [The Expression of Spontaneous Action in Japanese and Slovene]. *Slavistična revija* 41, 345–358.

Toporišič, J. (1982): *Nova slovenska skladnja [Syntax of Contemporary Slovene]*. Ljubljana: DZS.

Žele, A. (2001): *Vezljivost v slovenskem jeziku (s poudarkom na glagolu) [Valency in the Slovene Language (with Focus on the Verb)]*. Ljubljana: Založba ZRC, ZRC SAZU.

Žele, A. (2003a): *Glagolska vezljivost: iz teorije v slovar [Valency of Verbs: From Theory Towards Dictionary]*. Ljubljana: Založba ZRC, ZRC SAZU.

Žele, A. (2003b): Slovenska skladnja z vidika skladenjskih teorij [Slovene Syntax from the Point of View of Theories on Syntax]. *Slavistična revija* 51 (Posebna številka) [Special Issue], 141–163.

Kržišnik, E. (2004): Značilen tip minimalnih frazemov v slovenščini [A Distinctive Type of Phrasemes in Slovene].
<http://www-gewi.kfunigraz.ac.at/gralis/GraLiS%202004/Krzisnik%20Frazemy.htm>.

MULTEXT-East Project: <http://nl.ijs.si/ME/V3/>.

Specifications and Notation for MULTEXT-East Lexicon Encoding:
<http://nl.ijs.si/ME/V3/msd/>.

Slovene Dependency Treebank Project: <http://nl.ijs.si/sdt/>.

Summary

Površinskoskladenjsko označevanje prostih glagolskih morfemov predstavlja pri avtomatski skladenjski analizi jezika eno največjih zastranitev. Status predlogov, zaimkov oz. prostih morfemov ob glagolu je napovedljiv predvsem iz pomenskih lastnosti posameznega glagola in tudi iz konteksta ter pomenskih lastnosti udeležencev glagolskega dejanja, zato sistemsko označevanje (glede na tip struktur, v katerih se prostomorfemski glagoli pojavljajo, dejanja, ki jih takšni glagoli izražajo ipd.) za povsem natančno definiranje skladenjske vloge prostih morfemov sicer ni povsem primerno, je pa glede na trenutno razvojno fazo skladenjskega označevanja edino izvedljivo. Pravila za označevanje prostih glagolskih morfemov so zato zaenkrat oblikovana kot niz kompromisov, stopnja površinskoskladenjske poenostavitve pa je velika. Označevanja se bo torej treba lotiti postopno, v več fazah, računalniku pa bomo morali posredovati tudi pomenske podatke v slovarski obliki. Kljub temu pa pričakujemo, da bomo pri analizi korpusa Slovene Dependency Treebank o vzorcih pojavljanja določenega nabora skladenjskih struktur že zelo kmalu dobili zelo natančne podatke, ki slovenskemu jezikoslovju do zdaj še niso bili dostopni. K njegovemu razvoju lahko v trenutni fazi gradnje skladenjsko označenega korpusa pripomoremo že najmanj s tem, da opozorimo na številne strukture, ki zaenkrat še niso bile opisane.