Alexandr Rosen

Coping with Unruly Language: Non-Standard Usage in a Corpus

Abstract A language as used in real situations may differ substantially from its standard form. Before the entire range of NLP methods and tools can be applied to non-canonical variants of a language, appropriate categories for the analysis of deviant forms and constructions are needed, together with texts annotated by these categories. A discussion of non-standard language is followed by two case studies. The first study proposes a taxonomy of morphosyntactic categories as an attempt to analyze non-standard forms in non-native learners' Czech. The second study focuses on the role of a rule-based grammar and lexicon as tools for the detection and diagnostics of non-standard words and constructions in the process of building and using a parsebank.

Keywords Non-standard language, Czech, learner corpus, parsebank, treebank, constrain-based grammar, valency, HPSG

1 Introduction

In most cases, corpus annotation is not explicit about the canonicity of language use, although exceptions exist in specialized corpora or in specific cases in mainstream corpora (individual word forms – colloquial, dialectal or non-words). Non-standard usage defies general rules of grammar – it may involve performance errors, creative coinages, emerging phenomena. We start with the assumption that the text in a corpus and its linguistic annotation is where the two Saussurean faces of a single coin converge: the empirical evidence (language use, parole, performance, corpus) and the theory (language as a system, langue, competence, grammar). The annotation is also where multiple levels of analysis and linguistic theories may meet. An annotation scheme defined in terms of appropriate categories or even as a formal grammar can help to identify the difference between the regular and irregular, between the language as a system and its use.

It is often the case that instances of language use – in writing or speech of native and non-native speakers alike – do not comply with a norm or conventional pattern. The need to process non-standard language is growing, especially due to its ever more prominent presence in social media and the stepwise erosion of the role of language variants as social symbols or appropriate vehicles of communication, but also due to the increasing share of non-native speakers in many communities. The latter has additional consequences on the didactic front, represented mainly by the need to develop better methodologies suited to the non-native learner of a specific language.

Interestingly, linguistic variation impedes human communication only to a limited extent. Language users are able to recover meaning from idiosyncrasies on any level of the linguistic system and even recognize signals conveyed by the deviations to make guesses about the speaker's background or intention. On the other hand, standard NLP tools are usually much less adaptive and efficient when applied to non-standard language. Rule-based models, apparently vulnerable to any unexpected phenomena due to their dependence on (under-developed) conceptual categories and frameworks, are at a clear disadvantage. Stochastic models, generally more robust, seem to be in a better position. Possible strategies include applying a model trained on standard language, annotating more data, normalizing test data, deliberately corrupting training data, or adapting models to different domains. Eisenstein (2013) stresses the importance of a suitable match between the model and the domain of the text, while Plank (2016) points out that rather than to domains, the tools should be adapted to text varieties in a multi-dimensional space of factors such as dialect, topic, genre, gender, age, etc. Anyway, at least for rule-based or supervised models we lack suitable concepts and frameworks even distantly comparable to those for standard language. This leads us back to the issue of a suitable taxonomy and markup of unexpected phenomena – one of the topics of this paper (see section 3).

A rationalist approach to modeling non-standard language varieties has an important role not only in the design of categories suited for the analysis of non-standard forms and structures. Rather than being a random collection of unrelated phenomena, each variety represents a system, with rules and principles partially shared with other varieties, standard or non-standard. Deviations from the standard often represent regularly occurring patterns, such as spelling errors due to attraction in subject-predicate agreement.[1] There are many other regular phenomena which occur in the process of acquisition of non-native

1 A 100M corpus of Czech (SYN2010, see http://korpus.cz) includes 47 instances of short distance subject-predicate agreement patterns including spelling errors in masculine animate past tense forms, where the *-ly* ending is used instead of the correct homophonous *-li* ending (Dotlačil 2016).

language, some of them universal or specific to the target language, some of them due to the influence of the native or some other language already known to the learner. These deviations reveal facts about the speaker, her target and native language and can be used in methods and tools identifying the speaker and her background. Discovery of these rules and principles has practical benefits for foreign language teaching, forensic linguistics, the identification of the author's first language or the processing of non-standard language in general.[2]

A general discussion of issues related to non-standard language (section 2) is followed by two case studies. The first study (section 3) presents a taxonomy of learner language phenomena as an attempt to analyze non-standard forms produced by non-native speakers of Czech. The second study (section 4) focuses on the role of a rule-based grammar and lexicon as tools for the detection and diagnostics of non-standard words and constructions in the process of building and using a parsebank.

2 Non-standard language and its types

What counts as non-standard language? According to Bezuidenhout (2006), non-standard use of a language is one that "flouts a linguistic convention or that is an uncommon or novel use." The standard, conventional use is based on an explicit or implicit agreement among members of a linguistic community about the appropriate form of the language, given a specific situation.

This definition is problematic – it may not include some common language varieties that are quite far from the assumption about a standard, both in traditional linguistics or in NLP, such as Twitter messages. It might be useful to position specific varieties within a space of oppositions: the prescriptive or literary norm in contrast to colloquial, dialectal, 'uneducated' or archaic use; the language as a system (*langue*, the idealized linguistic competence) in contrast to the real use of language (*parole*, linguistic performance); written in contrast to spoken varieties; native in contrast to non-native language; the language of a child in contrast to the language of an adult native speaker; the language of people without language disorders in contrast to those with such handicaps; and also expectations of the grammar writer in contrast to anything else. Then we could delineate our notion of non-standard language to include varieties: (i) as used beyond the community of native speakers, (ii) of non-literary language (iii) of spoken language, and (iv) including deviations due to the specifics of language production, i.e. performance errors of all sorts.

2 E.g. typing assistants could offer an option to handle colloquial forms.

On the other hand, Hirschmann et al. (2007) define 'non-canonical' utterances in learner texts as:

> "[...] structures that cannot be described or generated by a given linguistic framework – canonicity can only be defined with respect to that framework. A structure may be non-canonical because it is ungrammatical, or it may be non-canonical because the given framework is not able to analyze it. For annotation purposes the reason for non-canonicity does not matter but for the interpretation of the non-canonical structures, it does. Most non-canonical structures in a learner corpus can be interpreted as errors [...] whereas many non-canonical structures in a corpus of spoken language or computer-mediated communication may be considered interesting features of those varieties."

This 'technical' view of what counts as non-standard language is more suitable to the tasks of annotating Czech as a foreign language and analyzing non-standard linguistic phenomena in a parsebank of Czech. After all, as Hirschmann et al. (2007) note, even if the interpretation of non-canonical structures differs for non-native and native speakers, many issues related to their appropriate annotation or analysis are shared.

Non-standard language can be detected, diagnosed and annotated by NLP methods in various ways (Meurers 2013; Meurers and Dickinson 2017). Tools developed for standard language and trained on standard or non-standard language can be applied (Ramasamy et al. 2015), texts can be manually annotated to build more task-specific models (Aharodnik et al. 2013), hand-crafted rules targeting relevant varieties can be used. It seems that designing an annotation scheme specific to non-standard language to build such a model brings better results (Berzak et al. 2016) than efforts to shoehorn existing annotation schemes to fit learner data (Cahill 2015). These results point to the need of "non-canonical categories for non-canonical data" (Dickinson and Ragheb 2015). Such categories are not part of common linguistic wisdom. It is not clear how to design a layered taxonomy of errors, an intelligibility metrics or a specification of the influence of other languages. The following section includes a proposal for a taxonomy of some phenomena of non-native Czech.

3 Designing categories for Czech as a foreign language

With the advance of learner corpora, the language produced by non-native speakers has been analyzed from perspectives familiar to corpus linguists but not so common in the field of language acquisition: learner texts are annotated by morphological and syntactic categories and structures, surveyed by statistical tools, and used to build stochastic models. Additional annotation, specific to learner language, has been used to capture non-standard phenomena: deviant forms and structures are assigned *target hypotheses* (corrections) and/or error types. So far, there are no standard solutions to these tasks.[3] Principles of emendation, error taxonomies and the shape of annotation schemes differ between projects, reflecting different answers to questions such as: What aspects of learner languages should be annotated? To what extent should the error taxonomy reflect standard linguistic categories and levels? Should multiple hypotheses be allowed, both in correction and error annotation? Is there any alternative to error annotation linked to a specific target hypothesis or can learner texts be analyzed and annotated as *interlanguage*, a language sui generis, approximating the target language in the process of language acquisition, to some extent independently of the target language?

A common strategy is to base the annotation on the concepts of native speakers' grammar, marking up deviations from the standard language in terms of errors in spelling, morphology, syntax, lexical choice, phraseology or register. However, some of the questions must be answered anyway: a nominal form, supposedly an object argument, marked by an incorrect morphological case, could be an error in spelling, morphology or syntax. An annotation scheme may insist on a single choice among these options or allow for their simultaneous specification as disjunctive hypotheses. Forms that do not match any existing word of the standard language (non-words, out-of-lexicon forms) present additional issues.

One possible starting point is a taxonomy of word classes based on a consistent partitioning along the morphological, syntactic and semantic criteria. These criteria are used as a mix in the definition of the standard sets of 8–10 word classes. For some of them, the three criteria yield the same result, but other classes are heterogeneous. A relative pronoun, defined by its semantic property of referentiality to an antecedent, may have an adjectival declension pattern as its morphological property, but it can be used in its syntactic role in a nominal

3 For examples of some tagsets used to annotate learner language see, e.g., http://merlin-platform.eu or https://www.linguistik.hu-berlin.de/de/institut/professuren/korpuslinguistik/forschung/falko.

position.[4] The class of Czech second position clitics consists of auxiliaries, weak pronouns or particles. Auxiliaries, prepositions and reflexive particles may be seen paradigmatically as parts of analytical paradigms in periphrastic verb forms, nouns in "prepositional cases", inherently reflexive verbs, while the rules of syntax treat the independent functional morphemes as individual syntactic words to make sure that they obey constraints on ordering, agreement or government. Thus, morphology, syntax and semantics take different perspectives, calling for a cross-classification of linguistic units at least along the three dimensions of morphology, syntax and semantics. It has been noted before (Díaz-Negrillo et al. 2010) that a cross-classifying scheme can be applied to texts produced by non-native learners. For English, the use of an adjective in an adverbial position can be analyzed as a mismatch between adverb as the syntactically appropriate category and adjective as the lexical category of the form used by the author of the text. A parallel Czech example is shown in (1), where the adjectival form *krásný* 'beautiful' is used instead of the standard adverbial form *krásně* 'beautifully'. The word can be annotated as a morphological adjective and syntactic adverb.

(1) Whitney Houston zpívala **krásný** → krásně
Whitney Houston sang beautiful → beautifully
'Whitney Houston sang beautifully.'

However, a morphologically rich interlanguage often deviates not just in the use of word classes but also in morphology. In (2), *táta* 'daddy' is nominative, but as the object of *viděl* 'saw' it should be accusative, which could be represented in the cross-classifying taxonomy as a mismatch between morphology and syntax in the category of case. A parallel example in English would be (3)[5] or, with a mismatch in number (4).

(2) Lucka viděla **táta** → tátu
Lucy.nom saw daddy.NOM → daddy.ACC
'Lucy saw her dad.'

(3) I must play with **he**.NOM → him.ACC

(4) The first year **have**.PL → has.SG been wonderful.

4 For a more detailed description of the proposed taxonomy of word classes see Rosen (2014).

5 The example is taken from Dickinson and Ragheb (2015).

In (5), the aspect of the content verb *napsat* 'to write' is perfective, while the auxiliary verb *bude* can only form an analytical future tense with an imperfective form. A perfective verb is used in its present form to express future meaning, as in (6).

(5) Eva **bude** **napsat** dopis
Eva will write.PFV letter
'Eva will write a letter.' (intended)

(6) Eva napíše dopis
Eva writes.PFV letter
'Eva will write a letter.'

Although the cross-classification idea can be applied to the analysis of all the above examples as mismatches between morphology and syntax, it does not seem to be the most intuitive solution. The annotation of (3) is agnostic about the fact that *he* is in a wrong case after all, a fact that should probably be avoided in the annotation of interlanguage, but which seems to be intuitive and important anyway. The form is only nominative rather than both nominative and accusative. While nominative is the morphological category, the missing syntactic interpretation is that of an object, a category specific to the layer of syntax.

The original proposal of Díaz-Negrillo et al. (2010) is concerned with English learner texts, assuming only standard POS labels at three layers: distribution (syntax), morphology and lexical stems. In standard language, the evidence from the three levels converges on a single POS. Mismatches indicate an error: stem vs. distribution (*they are very kind and* ***friendship***), stem vs. morphology (*television, radio are very* ***subjectives***), distribution vs. morphology (*the first year* ***have*** *been wonderful*). All these types are attested in Czech, but due to a wide range of phenomena related to morphonology and morphology, bare POS and mismatches of this type are not sufficient.

Our proposal combines error annotation with "linguistic" annotation of the original and the corrected version of the text, using standard categories such as domain-specific word class and other morphosyntactic properties as far as possible. Linguistic annotation of the original text may thus result in some forms labelled as unknown. Error annotation is based on the relation between the original and the corrected form, and on the relation between their analyses. An error is analyzed from three perspectives: (i) domain (see below), (ii) register (style), which is used as the benchmark to determine the error status, and (iii) location within the form, specified in terms of character positions and – if possible – in terms of a morpheme, such as stem, prefix, derivational suffix or inflectional ending. We propose five domains: spelling, morphonology, morphology, syntax and lexicon. Errors in each of the domains can be specified in more detail.

Spelling errors include word boundaries, punctuation, missing or incorrect capitalization (***mannheim*** →*Mannheim*), confusion of the homophonous vowels *i* and *y* (***lingvistyka*** →*lingvistika*), absence of graphemes such as *ě*, expressing palatalization of a preceding consonant (***ďeti*** → *děti*) or *j* followed by *e* as phonemes (***vjec*** → *věc*), and other issues connected with the use of diacritics.

Morphonology includes problems in palatalization, epenthesis or other processes, such as redundant presence or wrong absence of a vowel in some inflectional paradigms (***pesa*** → *psa* 'dog.ACC.SG' from *pes* 'dog.NOM.SG'; ***sestr*** → *sester* 'sister.GEN.PL' from *sestra* 'sister.NOM.SG'), incorrect presence or absence of vocalized versions of prepositions (***v*** *Vietnamu* → *ve Vietnamu* 'in Vietnam'), or confusion of voiced and devoiced consonants (***sůstala*** → *zůstala* 'stayed'). Given a target hypothesis, most errors in spelling and morphonology can be diagnosed automatically.[6]

Morphology includes paradigmatic errors related to inflectional patterns, including both non-words (*na* ***Erasmuse*** → *Erasmu* 'on the Erasmus'; ***studovám*** → *studuju* 'I study') and existing forms of the given word, inappropriate in the given context. If the original word exists, the error can be morphological or syntactic: *viděla* ***táta*** → *viděla tátu* '[she] saw [her] dad' (2).

Syntax covers syntagmatic issues: word order and incorrect use of word forms in a given context, including improper expression of valency, agreement, quantification etc.

Lexical errors typically concern the use of a semantically or syntactically inappropriate lexeme or even category such as verbal aspect, missing reflexive particle in inherently reflexive verbs, or an issue in phraseology.

It is often difficult to decide about the domain, i.e. about the cause of a specific deviation – is the issue in (2) an error in spelling, morphology or syntax? One possible strategy is to apply a rule selecting a single option. In the manual annotation of the *CzeSL* corpus (Rosen et al. 2014), the rule was to specify the deviation in a domain where the analysis requires a more sophisticated judgment, e.g. morphology or syntax in preference to spelling. An alternative strategy is to specify the deviation in parallel in all relevant domains. This solution leaves the decision open for additional analysis and fits well in the concept of cross-classification.

The combined error and linguistic annotation can be used to tag the corpus and to specify types located within a hierarchy of learner language phenomena. The error annotation together with the two poles of linguistic annotation – one for the ill-formed and one for the corrected word – represent a pattern. For a

6 See Jelínek et al. (2012) for a list of "formal errors": missing or redundant character, character metathesis, etc., which can often be interpreted in linguistic terms.

simple case such as (2), the pattern is shown in Table 1.[7] A taxonomy of such patterns can be built, and references to more or less abstract patterns can be used as tags. A more abstract pattern in Table 2 represents all cases where a nominative form is used instead of an accusative form.

Table 1: The pattern for *táta* in (2) (*Lucka viděla* ***táta*** → *tátu* 'Lucy saw her Dad').

		error annotation	linguistic annotation	
			original	target
location		inflectional suffix	–	–
register		standard	–	–
domain	spelling	character replacement	a	u
	morphology	case	nominative	accusative
	syntax	valency	object of *viděla*	object of *viděla*

Table 2: The abstract pattern for a form which is nominative instead of accusative.

		error annotation	linguistic annotation	
			original	target
location		inflectional suffix	–	–
register		standard	–	–
domain	morphology	case	nominative	accusative

A different type of error is shown in (7). Unlike *táta* in (3), *babičkem* is a nonword. However, it can be interpreted as consisting of the feminine stem *babičk-* and the masculine singular instrumental suffix *-em*, compatible with the preposition but incompatible with the gender of the stem.[8]

(7) Byl jsem doma s **babičkem** → babičkou
was AUX at home with granny(F).M.SG.INS granny(F).F.SG.INS
'I was at home with Grannie.'

The pattern is shown in Table 3. A more abstract pattern could include only the location and morphology rows.

7 In a fully specified pattern, morphological analysis concerns all relevant categories, including lemma.

8 The bare suffix is ambiguous. It can also express present tense first person plural of some verbal paradigms (*nesem* '[we] carry'). Rather than suggesting such unlikely alternatives, the author is given the benefit of the doubt. For the same reason, we refrain from hypothesizing 'grandpa' (*s dědečkem*) rather than 'granny' (*s babičkou*).

Table 3: The pattern for *babičkem* in (7).

		error annotation	linguistic annotation	
			original	target
location		inflectional suffix	–	–
register		standard	–	–
domain	spelling	two characters' replacement	*em*	*ou*
	morphology	stem/suffix mismatch	stem feminine, suffix masculine	

Tags referring to such patterns can be used as a powerful indicator of the type of interlanguage and the language learner's competence, and can help to build models of interlanguage by machine learning methods. The scheme will be evaluated in trial annotation, including inter-annotator agreement, and tested in machine learning experiments.

Manual annotation can be supported or even replaced by automatic identification of some error types (Jelínek et al. 2012), coupled with a tool suggesting corrections (Ramasamy et al. 2015). Some annotation of a learner corpus can thus be done automatically, without the involvement of human annotators in the process (Rosen 2017).

4 Identifying non-standard language in a corpus

Annotation of word forms and structures in a corpus rarely distinguishes standard language from other varieties. Except for individual word forms in mainstream corpora and error annotation in learner corpora, systematic accounts of non-standard usage are virtually missing. In addition to colloquial, dialectal, obsolete and bookish expressions or imports, described in available lexical resources, non-standard language may also involve performance errors, creative coinages, or emerging phenomena. Most of these phenomena are not covered by standard grammars, but they are still not random, even though the underlying patterns are not easy to discover. In this section, we show an attempt to detect and annotate these phenomena in a treebank/parsebank of Czech.

The theoretical assumption is that linguistic annotation of a corpus represents the meeting point of the empirical evidence (parole) and the theory (langue), in the sense of Saussurean *sign* (de Saussure 1916). Moreover, the annotation is also where multiple levels of analysis and linguistic theories may meet and be explicit about any, even irregular, phenomena. An annotation scheme defined as a formal grammar can help to identify the difference between the regular and irregular, between the language as a system and the use of language.

This is the motivation behind the project of a corpus annotated by standard stochastic tools[9] and checked by a rule-based grammar and valency lexicon, which are also used to infer additional linguistic information about the annotated data.[10] The grammar has the role of a watchdog: to check stochastic parses for both formal and linguistic correctness and consistency. Compliant parses receive additional information: lexical categories receive valency frames to be saturated by complements and project relevant properties to phrasal nodes. Ideally, the grammar should define standard language in the sense of Hirschmann et al. (2007, see section 2 above), although in real life the grammar both overgenerates, leaving some non-standard utterances undetected, and undergenerates, deciding that some standard utterances are not correct.

The grammar consists of a lexical module, providing valency frames, and a syntactic module, checking the parse and projecting information in lexical heads to phrases and complements (dependents). The lexical module, operating on lexical entries derived from external valency lexica, generates available diatheses. The syntactic module matches the generated lexical entries with the data. Categorial information about words and phrases in the data and the lexicon is structured according to a cross-classifying taxonomy, capturing all distinctions present in the standard Czech tagset used in the stochastic parse.[11]

The grammar is implemented in Trale,[12] a formalism designed for grammars based on HPSG, a linguistic theory modeling linguistic expressions as typed feature structures.[13] The grammar differs from a standard implemented HPSG grammar mainly in its role of a constraint solver, rather than a parser or generator. The constraints come from three sources: data, lexicon, and grammar proper. No syntactic rules of the context-free type are needed because the grammar operates on structures already built by a stochastic parser – the syntactic backbone is present in the data, where each sentence has a single parse. Ambiguities or underspecifications may arise only due to the more detailed taxonomy in the treebank format and/or an uncertainty about the choice of a valency frame.

9 See Jelínek (2016).

10 For more detail about the project see, e.g., Petkevič et al. (2015a).

11 See also Petkevič et al. (2015b) for a description of the annotation of periphrastic verb forms using an additional analytical dimension. Periphrastic verb forms are treated with respect to their dual status, i.e. from the paradigmatic perspective as forms of the content verb, and from the syntagmatic perspective as constructions.

12 http://www.ale.cs.toronto.edu/docs/

13 See, e.g., Pollard and Sag (1994) or Levine and Meurers (2006).

The lexical module uses two external valency lexicons: VALLEX[14] and PDT-VALLEX,[15] with their deep valency frames and information about the forms of the syntactic arguments (case, verbal form, etc.). The frames reflect the Praguian valency theory of the Functional Generative Description (Panevová 1994). The lexical module provides the mapping of the frames to their instantiations in specific verbal diatheses and morphological forms, using the same formalism as the syntactic component.

If the syntactic module, after checking the parse using the lexical specifications, decides that the parse complies in all respects, the structure is provided with all available information. If, however, some predicates are left without valency frames, completeness and coherence of the argument structure cannot be checked. Yet some phenomena, such as grammatical agreement, can still be checked. A failure can also be caused by a valency frame. If so, the sentence is additionally checked without that frame. A sentence may also fail due to constraints of the syntactic module. Then the last and weakest test is applied, using only the data format definition without constraints.

Any of these checks may fail due to non-standard linguistic phenomenon in the data, an incorrect decision of the parser or the tagger, or an error in the grammar or lexicon. An efficient and powerful diagnostic is an important task for the future. One option is to make use of the constraint-based architecture by successively relaxing constraints to find the grammatical or lexical constraint and the part of the input responsible for the failure. Another possibility is to use constraints targeting specific non-standard structures or lexical specifications.[16]

Non-standard phenomena can be detected precisely because a grammar of linguistic competence can never fit the corpus as the evidence of linguistic performance completely. To distinguish the cases of truly non-standard language from problems of the grammar on the one hand and to identify and diagnose the types of non-standard language on the other, the diagnostics should be extended to find which specific constraints are violated by which specific words or constructions in the data.

The examples below illustrate the role of the grammar. In (8) and (9) the possessive form agrees in gender and case (and number) with the head noun.

14 See http://ufal.mff.cuni.cz/vallex, Lopatková et al. (2008), Žabokrtský and Lopatková (2007).

15 See Hajič et al. (2003).

16 The so-called *mal-rules* have been used in the context of CALL (computer-assisted language learning) at least by Schneider and McCoy (1998, for users of American Sign Language learning English as their L2), Bender et al. (2004), and Flickinger and Yu (2013) – both implemented in HPSG.

Examples (10) and (11) are different: in (10) the possessive form does not agree with the head noun either in case or in gender, in (11) both in case and gender. Note that the possessive form in (10), which is the same as in (8), does not strike many speakers as incorrect. In the SYN2015 corpus, the share of these non-standard forms is about 4% in the total number of masculine dative singular NPs preceded by the preposition *k*. Example (11) has a similar status, but it is acceptable only to speakers of a dialect of Czech.

(8) Přitiskl se k otcově noze
clung REFL to father's.F.DAT leg(F).DAT
'He pressed against his father's leg.'

(9) Přistoupil k otcovu stolu
approached to father's.M.DAT table(M).DAT
'He appoached his father's table.'

(10) Přistoupil k ?**otcově** stolu
approached to father's.M.LOC/F.DAT table(M).DAT
'He appoached his father's table.'

(11) Přistoupil k ?**otcovo** stolu
approached to father's.N.NOM/ACC table(M).DAT
'He appoached his father's table.'

While (10) and (11) could be seen as examples of suboptimal morphology, (12)–(15) show suboptimal syntax. In (12), an example of *zeugma*, the two coordinated verbs are supposed to share a single object. However, the form of the object (a prepositional phrase) is consistent only with the second verb. In (13), the position of the indirect object of the matrix clause is filled twice: by the headless relative clause and by the personal pronoun. In a standard structure, only a headed relative clause is compatible with an indirect object in the dative case (14). Finally, the matrix clause in (15) includes a subject of the embedded clause (*Gazda*).

(12) ?? Včera jsem viděl a mluvil **s tím člověkem**
yesterday AUX saw and talked with that man
'Yesterday I saw and talked to that man.'

(13) ? **Kdo přijde pozdě**, nic mu nedají
who.nom comes late nothing.ACC him.DAT NEG. give.3.PL
'Who comes late won't get anything.' (intended)

(14)	Tomu,	kdo	přijde	pozdě,	nic	nedají
	that.DAT	who.NOM	comes	late	nothing.ACC	NEG.give.3.PL
	'Who comes late won't get anything.'					

(15) ??	Nebo	já	**Gazda**	nevím,	jak	diktuje
	or	I	Gazda	NEG.know.1.SG	how	dictates
	'Or I don't know how Gazda dictates.'					

In most of the above examples, the stochastic parser ignores the agreement mismatch or the structural anomaly and builds a correct tree. On the other hand, the grammar does not accept the parse, which is the required result. Like every rule-based grammar, it has limited coverage, but a missing account of a phenomenon only means that the grammar overgenerates (is too permissive). Filling gaps in the coverage is another priority for the future.

The grammar and lexicon have been developed and tested on a set of 876 sentences, extracted from the annotation manual of the Prague Dependency Treebank (Hajič et al. 1997), representing a wide range of linguistic phenomena. For 592 sentences a valency frame from the lexicon was found. The number of sentences verified by the grammar is 560. This includes 301 sentences with a valency frame. For more extensive testing, the SYN2015 corpus was used, including about 100 million words, i.e. 7.2 million sentences. For 77% of sentences, at least one valency frame was found and 55% of sentences passed the grammar, 16% including a valency frame, 23% without any valency frame, and 16% after the valency frame was dropped. The next step is to categorize the failures and build a corpus showing the results, including the grammar flags, in a user-friendly way.

5 Conclusion

We have presented two ways to approach non-standard language, with a stress on its proper detection and diagnosis. In the design of an annotation scheme for Czech of non-native learners, we have shown an approach to the analysis of non-standard word forms and structures, based on a layered description of the original and the target expression, combined with corresponding error annotation. In the second study, a method was presented for the detection and diagnosis of non-standard forms and expressions in the grammar-checked annotation of a parsebank. We see this effort as an attempt to tackle a domain of growing importance, one in which the methods and tools available for standard language have only limited usability. Admittedly, we have merely scratched the surface of the topic.

Acknowledgements

This work has been supported by the Grant Agency of the Czech Republic, grant no. 16-10185S.

References

Aharodnik, Katsiaryna, Marco Chang, Anna Feldman and Jirka Hana. 2013. Automatic identification of learners' language background based on their writing in Czech. In *Proceedings of the 6th International Joint Conference on Natural Language Processing (IJNCLP 2013), Nagoya, Japan, October 2013*, 1428–1436.

Bender, Emily M., Dan Flickinger, Stephan Oepen, Annemarie Walsh and Tim Baldwin. 2004. Arboretum: Using a precision grammar for grammar checking in CALL. In *Proceedings of the InSTIL/ICALL Symposium: NLP and Speech Technologies in Advanced Language Learning Systems*, Venice, Italy.

Berzak, Yevgeni, Jessica Kenney, Carolyn Spadine, Jing Xian Wang, Lucia Lam, Keiko Sophie Mori, Sebastian Garza and Boris Katz. 2016. Universal dependencies for learner English. *CoRR*, abs/1605.04278.

Bezuidenhout, Anne L. 2006. Nonstandard language use. In Keith Brown (ed.) *Encyclopedia of Language and Linguistics. Second Edition*, 686–689. Oxford: Elsevier.

Cahill, Aoife. 2015. Parsing learner text: to shoehorn or not to shoehorn. In *Proceedings of The 9th Linguistic Annotation Workshop*, 144–147, Denver, Colorado, USA. Association for Computational Linguistics.

Chomsky, Noam. 1970. Remarks on nominalization. In R. Jacobs and P. Rosenbaum (eds.), *Reading in English Transformational Grammar*, 184–221. Waltham: Ginn and Co.

Dickinson, Markus and Marwa Ragheb. 2015. On grammaticality in the syntactic annotation of learner language. In *Proceedings of The 9th Linguistic Annotation Workshop*, 158–167. Denver, CO.

Dotlačil, Jakub. 2016. Shoda podmětu s přísudkem, pravopis a iluze gramatičnosti. A talk presented at the conference Linguistics and Literary Studies: Paths and Perspectives, Liberec, 22–23 September 2016.

Díaz-Negrillo, Ana, Detmar Meurers, Salvador Valera and Holger Wunsch. 2010. Towards interlanguage POS annotation for effective learner corpora in SLA and FLT. *Language Forum*, 36(1–2):139–154. Special Issue on Corpus Linguistics for Teaching and Learning. In Honour of John Sinclair.

Eisenstein, Jacob. 2013. What to do about bad language on the internet. In *Proceedings of the North American Chapter of the Association for Computational*

Linguistics (NAACL), 359–369, Stroudsburg, Pennsylvania. Association for Computational Linguistics.

Flickinger, Dan and Jiye Yu. 2013. Toward more precision in correction of grammatical errors. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task*, 68–73, Sofia, Bulgaria. Association for Computational Linguistics.

Hajič, Jan, Jarmila Panevová, Eva Buráňová, Zdenka Urešová and Alla Bémová. 1997. A manual for analytic layer tagging of the Prague Dependency Treebank. Technical Report TR-1997-03, ÚFAL MFF UK, Prague, Czech Republic.

Hajič, Jan, Jarmila Panevová, Zdeňka Urešová, Alevtina Bémová and Petr Pajas. 2003. PDT-VALLEX: Creating a large-coverage valency lexicon for treebank annotation. In *Proceedings of The Second Workshop on Treebanks and Linguistic Theories*, 57–68. Växjö University Press.

Hirschmann, Hagen, Seanna Doolittle and Anke Lüdeling. 2007. Syntactic annotation of non-canonical linguistics structures. In *Proceedings of Corpus Linguistics 2007*, Birmingham.

Jelínek, Tomáš. 2016. Combining dependency parsers using error rates. In *Text, Speech and Dialogue – Proceedings of the 19th International Conference TSD 2016*, 82–92. Springer.

Jelínek, Tomáš, Barbora Štindlová, Alexandr Rosen and Jirka Hana. 2012. Combining manual and automatic annotation of a learner corpus. In Petr Sojka, Aleš Horák, Ivan Kopeček and Karle Pala, editors, *Text, Speech and Dialogue: 15th International Conference*, 127–134. Berlin/Heidelberg: Springer.

Levine, Robert D. and Walt Detmar Meurers. 2006. Head-Driven Phrase Structure Grammar: Linguistic approach, formal foundations, and computational realization. In Keith Brown (ed), *Encyclopedia of Language and Linguistics. Second Edition.* Oxford: Elsevier.

Lopatková, Markéta, Zdeněk Žabokrtský and Václava Kettnerová. 2008. *Valenční slovník českých sloves.* Praha: Karolinum.

Meurers, Detmar. 2013. Natural language processing and language learning. In C. A. Chapelle (ed.), *Encyclopedia of Applied Linguistics*, 4193–4205. Blackwell.

Meurers, Detmar and Markus Dickinson. 2017. Evidence and interpretation in language learning research: Opportunities for collaboration with computational linguistics. *Language Learning, Special Issue on Language learning research at the intersection of experimental, corpus-based and computational methods: Evidence and Interpretation.* Language Learning, 67(S1): 66–95.

Panevová, Jarmila. 1994. Valency frames and the meaning of the sentence. In P. A. Luelsdorff (ed.), *The Prague School of structural and functional linguistics. A short introduction*, 223–243. Amsterdam/Philadelphia: John Benjamins.

Petkevič, Vladimír, Alexandr Rosen and Hana Skoumalová. 2015a. The grammarian is opening a treebank account. *Prace Filologiczne*, LXVII: 239–260.

Petkevič, Vladimír, Alexandr Rosen, Hana Skoumalová and Přemysl Vítovec. 2015b. Analytic morphology – merging the paradigmatic and syntagmatic perspective in a treebank. In Jakub Piskorski, Lidia Pivovarova, Jan Šnajder, Hristo Tanev and Roman Yangarber (eds.), *The 5th Workshop on Balto-Slavic Natural Language Processing (BSNLP 2015)*, 9–16, Bulgaria: Hissar.

Plank, Barbara. 2016. What to do about non-standard (or non-canonical) language in NLP. In *KONVENS 2016*.

Plank, Barbara, Hector Martinez Alonso and Anders Søgaard. 2015. Non-canonical language is not harder to annotate than canonical language. In *The 9th Linguistic Annotation Workshop (held in conjunction with NAACL 2015)*, 148–151. Association for Computational Linguistics.

Pollard, Carl J. and Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. Chicago: University of Chicago Press.

Ramasamy, Loganathan, Alexandr Rosen, and Pavel Straňák. 2015. Improvements to Korektor: A case study with native and non-native Czech. In Jakub Yaghob, editor, *ITAT 2015: Information technologies – Applications and Theory / SloNLP 2015*, 73–80, Charles University in Prague.

Rosen, Alexandr. 2014. A 3D taxonomy of word classes at work. In Ludmila Veselovská and Markéta Janebová, editors, *Complex Visibles Out There. Proceedings of the Olomouc Linguistics Colloquium 2014: Language Use and Linguistic Structure*, volume 4 of *Olomouc Modern Language Series*, 575–590. Olomouc: Palacký University.

Rosen, Alexandr. 2017. Introducing a corpus of non-native Czech with automatic annotation. In Piotr Pęzik, Jacek Waliński, and Krzysztof Kosecki, editors, *Language, Corpora and Cognition*, 163–180. Frankfurt am Main, Bern, Bruxelles, New York, Oxford, Warszawa, Wien: Peter Lang.

Rosen, Alexandr, Jirka Hana, Barbora Štindlová, and Anna Feldman. 2014. Evaluating and automating the annotation of a learner corpus. *Language Resources and Evaluation – Special Issue: Resources for language learning*, 48(1): 65–92.

de Saussure, Ferdinand. 1916. *Cours de linguistique générale*. Paris. Publié par Ch. Bally et A. Sechehay avec la collaboration de A Riedlinger.

Schneider, David and Kathleen F. McCoy. 1998. Recognizing syntactic errors in the writing of second language learners. In Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics Volume 2, ACL '98, 1198–1204, Stroudsburg, PA, USA. Association for Computational Linguistics.

Žabokrtský, Zdeněk and Markéta Lopatková. 2007. Valency information in VALLEX 2.0: Logical structure of the lexicon. *The Prague Bulletin of Mathematical Linguistics*, (87): 41–60.