Modeling non-standard language

Alexandr Rosen
Institute of Theoretical and Computational Linguistics
Faculty of Arts, Charles University
Prague, Czech Republic
alexandr.rosen@ff.cuni.cz

Abstract

A specific language as used by different speakers and in different situations has a number of more or less distant varieties. Extending the notion of non-standard language to varieties that do not fit an explicitly or implicitly assumed norm or pattern, we look for methods and tools that could be applied to such texts. The needs start from the theoretical side: categories usable for the analysis of non-standard language are not readily available. However, it is not easy to find methods and tools required for its detection and diagnostics either. A general discussion of issues related to non-standard language is followed by two case studies. The first study presents a taxonomy of morphosyntactic categories as an attempt to analyse non-standard forms produced by non-native learners of Czech. The second study focusses on the role of a rule-based grammar and lexicon in the process of building and using a parsebank.

1 Introduction

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 conventional pattern specified in standard handbooks or seen as appropriate by the community of native speakers.[1] Yet the message is often intelligible and the communication is not hampered by linguistic variations. Language users are able to recover meaning from idiosyncrasies on any level of the linguistic system, such as phonetics, phonology, graphemics, morphology, syntax, semantics or pragmatics, including peculiarities occurring on multiple levels in parallel. In addition to understanding the content of the message, the hearer is often able to recognize implicit signals conveyed by any deviations from the expected and appropriate *register* and may even use various linguistic signs to make guesses about the speaker's background or covert intention.

Such abilities of the language user are in sharp contrast with the rigidity and performance of most language models. While rule-based models are very vulnerable to any unexpected phenomena and appropriate categories usable for their analysis are not readily available, stochastic models seem to be in a better position to cope with non-standard language. Apart from being more robust in general, perhaps at the cost of lower precision, various strategies can be used instead of a naively applying a model trained on 'standard' language. Reviewing a range of options, such as 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 we should adapt our tools to text varieties in a multi-dimensional space of factors such as dialect, topic, genre, the speaker's gender, age, etc. Such models should be built using non-standard language as the training data to handle similar input. To handle code-switching and a mix of language varieties within a single text, multiple models may be needed in parallel. Alternatively, a single model can be trained on an appropriately annotated text as one of the 'domain adaptation' methods, which leads us back to the issue of a suitable taxonomy and markup of unexpected phenomena – one of the topics of this paper (see §3).

[1] The appropriateness is to a large extent determined by all sorts of contextual factors and the community may include only some privileged or elitist groups.

Proceedings of the Workshop on Grammar and Lexicon: Interactions and Interfaces,
pages 120–131, Osaka, Japan, December 11 2016.

We start with the assumption that there is an important role for a rationalist approach to language modeling in general, and to modeling of non-standard language varieties in particular. At the core of this premise is an observation that the varieties are not just random collection of unrelated phenomena, that each variety represents a system, with rules and principles partially shared with other varieties, standard or non-standard. In addition to its theoretical merit, the discovery of these rules and principles has practical benefits for many tasks in today's increasingly multilingual and globalized linguistic communities. These benefits include applications in the field of foreign language teaching, forensic linguistics, identification of the author's first language or processing of non-standard language in general for purposes of all sorts, better suited to the needs of both native and non-native speakers.[2]

We are aware of the wealth of studies targeting linguistic variability, including language development and acquisition, dialects, ethnolects, idiolects, specifics of spoken and colloquial language, and language disorders, in sociolinguistics and other fields. However, our specific aim is to explore options for extending the space of linguistic phenomena covered by existing language models beyond the limits of a standard language.

A general discussion of issues related to non-standard language (§2) is followed by two case studies. The first study (§3) presents a taxonomy of morphosyntactic categories as an attempt to analyse non-standard forms produced by non-native learners of Czech. The second study (§4) focusses on the role of a rule-based grammar and lexicon in the process of building and using a parsebank. Both topics are very partial probes into the general topic of non-standard language, but at least they target different issues, which can make the overall picture less patchy and more relevant.

2 Types of 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, because it may not include some common language varieties that are quite far from the standard use of a language, assumed both in traditional linguistics or in NLP, such as Twitter messages.[3] Rather than using the notion of standard as a universal yardstick, a more realistic view could be a point of reference relative to a binary opposition. It can be 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.

Most deviations from any of the above "standards" are not random. Representative corpora of native written language show that there are regularly occurring patterns of non-standard usage, such as orthographical errors due to attraction in subject-predicate agreement.[4] There are many other regular phenomena occurring in the process of acquisition of non-native 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, the target language and the native language and can be used in methods and tools identifying the language users and their background.

[2] A trivial example is represented by Czech typing assistants on portable devices. To the best of our knowledge, they do not offer any alternative to predicting standard word forms, ignoring any user preferences.

[3] While we do not agree with Plank et al. (2015) that the annotation of 'non-canonical' language is as hard (or as easy) as the annotation of newswire texts, we agree that "standard language" may be very different from what these traditional resources offer.

[4] According to Dotlačil (2016), SYN2010, the 100M Czech corpus (available at `http://korpus.cz`) includes 47 instances of short distance subject-predicate agreement patterns including purely orthographical errors in masculine animate past tense forms, where the *-ly* ending is used instead the correct homophonous *-li* ending.

A more practically oriented definition is offered by Hirschmann et al. (2007) in the context of annotating a learner corpus, referring to non-standard ('non-canonical') utterances 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 analyse 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 we cover in the present paper: annotating Czech as a foreign language and analyzing 'non-standard' linguistic phenomena in a parsebank of Czech. 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 by both tasks. However, we still feel the need to delineate the notion of non-standard language used here to include language varieties: (i) as used beyond the community of native speakers, (ii) of non-literary language, often widespread and representing a standard of its own kind, such as "Common Czech" (Sgall and Hronek, 1992), (iii) of spoken language, and (iv) including deviations due to the specifics of language production, i.e. performance errors of all sorts.

There are multiple ways how non-standard language can be processed, detected and diagnosed. As for learner texts, 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 (as it happens in learner corpora) and used to built stochastic models (Aharodnik et al., 2013), hand-crafted rules targeting standard and non-standard varieties can be used. While it is still true that "domain adaptation for parsing the web is still an unsolved problem" (Petrov and McDonald, 2012), it seems that designing an annotation scheme specific to non-standard (learner) language in order 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 which linguistic categories are suitable for the annotation of a non-standard text to design a tagset describing deviant word forms and syntactic structures, a taxonomy of errors at multiple levels of interpretation, an intelligibility metrics or even a specification of the influence of other languages.

The following section includes a proposal for a taxonomy of word forms, non-standard from the morphological perspective.

3 Designing categories for the morphology of Czech as a foreign language

With the increasing role of linguistic corpora, statistics and other formal analytical tools used in the field of language acquisition, demand is growing for categories applicable beyond standard language to the specific needs of the analysis of a language produced by non-native speakers. But before proceeding to the topic of a taxonomy suitable for the task of annotating such texts, we show some options of how standard assumptions about word classes could be modified. The resulting picture is actually a good starting point for an extension to non-standard language.[5]

Taxonomies of linguistic units such as morphemes, words, multiword expressions, phrases, clauses or sentences and of their properties are of critical importance to both theoretical and applied linguistics. Categories representing those units are crucial components in all rule-based and most stochastic models. The standard sets of 8–10 word classes (POS) are defined by a mix of morphological, syntactic and semantic criteria. For some POS the three criteria yield the same result, but POS such as numerals and pronouns end up as heterogeneous classes. A relative pronoun, defined by its semantic property of

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

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 position.

More evidence of multiple class membership is easy to find. In Czech, the second position clitic is a category that must be lexically specified as such, but it is an auxiliary, a weak pronoun or a particle at the same time. Auxiliaries, prepositions and reflexive particles are sometimes treated as internal to a single analytical paradigm: a periphrastic verb form, a noun in "prepositional case", or inherently reflexive verb, while the rules of syntax need to access 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. Unsurprisingly, the option of cross-classification is often mentioned in literature, but it is hardly ever pursued. One of the criteria is usually adopted as the main one and others as complementary. Semantics is favored e.g. by Brøndal (1928), morphology by Saloni and Świdziński (1985, p. 95), syntax by Grzegorczykowa et al. (1998, p. 59). In theoretical linguistics, the syntactic criterion prevails: four basic lexical categories, determined by the combinations of two binary features (Chomsky, 1970), correspond to labels in a syntactic tree. The syntactic perspective is even more explicit in Jackendoff (1977, p. 31–32), or Dechaine (1993) – see Table 1. The binary features can be used to specify hyperclasses, such as −nominal for verbs and prepositions, which both assign case. However, none of the feature systems in the table is able to capture classes distinguished by all relevant properties.

	Chomsky (1970)		Jackendoff (1977)		Dechaine (1993)	
	nominal	verbal	subject	object	referential	object
Nouns	+	−	+	−	+	−
Verbs	−	+	+	+	+	+
Adjectives	+	+	−	−	−	−
Adpositions	−	−	−	+	−	+

Table 1: A syntax-based taxonomy – features determining basic lexical categories

The morphology-based classification of Saloni and Świdziński (1985), based on the presence of specific inflectional categories as properties of a POS, shows how POS correlate with sets of morphological categories. However, a single item can have more than one set of such categories, as in the Czech example (1). Like personal pronouns, possessive pronouns are marked for (i) person, number and gender to agree with their antecedents and – like adjectives – for (ii) number, gender, case to agree with the modified noun. Cross-classification allows for the former set to be appropriate for pronouns as a semantic POS, while the former set represents the properties of morphological adjectives. Czech possessive pronouns belong to both classes at the same time.

(1) Jana přišla, ale jejího syna jsem neviděl.
Jana$^{\text{FEM,NOM}}$ came but her$^{\text{FEM,3RD}}_{\text{MASC,ACC}}$ son$_{\text{MASC,ACC}}$ I haven't seen
'Jana has arrived, but I haven't seen her son.'

The cross-classification approach has been proposed e.g. by Brøndal (1928) and Komárek (1999), but rarely presented in standard reference books. To handle gerunds and other hybrid categories, Lapointe (1999) proposes dual lexical categories, determining both the external and internal syntactic properties of the item. Similarly as Combinatory Categorial Grammar (Steedman and Baldridge, 2011), this approach points to cross-classification. HPSG (Pollard and Sag, 1994) goes a step further by representing words and phrases as objects consisting of the unit's morphological, syntactic and semantic properties. The individual properties may be used as interfaces to external theories or processing modules.

To model variations in non-standard language, occurring at multiple levels of the language system, cross-classification, a multi-dimensional or multi-level system seems to be a natural choice. In the rest of this section, we will focus on the application of multidimensional taxonomy to the language of non-native speakers. The primary focus is on Czech, but most points should be relevant also to other morphologi-

cally rich languages.

It has been noted before (Díaz-Negrillo et al., 2010) that a cross-classifying scheme can be usefully applied to texts produced by foreign language learners. The scheme treats such texts as specimens of *interlanguage*, a language sui generis, approximating the target language in the process of language acquisition, to some extent independently of the target language, i.e. of the error-based approach (Corder, 1981; Selinker, 1983). The non-standard features of interlanguage can be modelled as deviations on appropriate levels.

For English, the use of an adjective in an adverbial position can be analysed 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 (2), where the adjectival form *krásný* 'beautiful' is used instead of the standard adverbial form *krásně* 'beautifully'. The word can be annotated as morphological adjective and syntactic adverb.

(2) Whitney Houston zpívala ***krásný** → krásně.
Whitney Houston sang ***beautiful**$_{\mathbf{ADJ}}$ → beautifully$_{\mathrm{ADV}}$
'Whitney Houston sang beautifully.'

However, in Czech as a morphologically rich language, interlanguage typically deviates not just in the use of word classes, but also in morphemics, morphology and morphosyntax. A richer taxonomy is required that the one proposed in Díaz-Negrillo et al. (2010) for English. First of all, categories such as number, gender, case are needed. In (3), *táta* 'daddy' is nominative, but its syntactic role as the object of *viděl* 'saw' requires the accusative. This represents a mismatch between morphology and syntax in the category of case. A parallel example in English would be would be (4-a)[6] or, with a mismatch in number, (4-b).

(3) Lucka viděla ***táta** → tátu.
Lucy$_{\mathrm{NOM}}$ saw daddy$_{*\mathbf{NOM}\,\rightarrow\,\mathrm{ACC}}$
'Lucy saw her dad.'

(4) a. I must play with ***he**$_{\mathrm{NOM}}$ → him$_{\mathrm{ACC}}$.
b. The first year ***have**$_{\mathrm{PL}}$ → has$_{\mathrm{SG}}$ been wonderful.

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

(5) a. Eva bude ***napsat** dopis.
Eva will write$_{*\mathbf{PERF}}$ letter
'Eva will write a letter.'
b. Eva napíše dopis.
Eva writes$_{\mathrm{PERF}}$ letter.
'Eva will write a letter.'

Although the cross-classification idea can be applied to the analysis of all of the above examples as mismatches between morphology and syntax, it does not seem to be the most intuitive solution.

As Dickinson and Ragheb (2015) say: "While errors (i.e., ungrammaticalities) can be derived from mismatches between annotation layers, they are not primary entities. The multi-layer linguistic annotation is primarily based on linguistic evidence, not a sentence's correctness." Indeed, the annotation of (4-a) may be seen as agnostic about the fact that *he* is in a wrong case and that the accusative case can be accepted as a syntactic category. As the authors say: "the word *he* cannot simply be marked as a nominative or accusative pronoun because in some sense it is both. Thus, one may want to annotate multiple layers, in this case one POS layer for morphological evidence and one for syntactic distributional evidence (i.e., position)."

[6]The example is taken from Dickinson and Ragheb (2015).

Yet we see as legitimate a different claim, namely that 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. Moreover, it is not obvious that considerations related to correctness are absent from the analysis or secondary. We prefer to see the mismatch between annotation layers on the one hand and the aspect of comparison to the target (correct) form on the other as complementary.[7]

This modification of the original cross-classifying scheme is supported by more evidence from the domain of morphology. 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 of these types are attested in Czech, but due to a wide range of phenomena related to morphemics and morphology, bare POS and mismatches of this type are not sufficient.

To accommodate possibly parallel deviations in orthography, morphemics and morphology the number of layers must be extended, each with categories of its own. We start from an existing taxonomy for Czech learner texts with less granular morphological categories (Rosen et al., 2014), using the following layers to analyse non-standard forms, abstracting from other issues of syntax, semantics or pragmatics. Each of the layers is specified by its relevant linguistic category (stem, paradigm, case, number, gender, etc.) and possibly by an error label. The first two items are actually not layers in the linguistic sense but rather specifications from a different viewpoint.

- Formal: missing or redundant character, character metathesis, etc.
- Location: identification of the part of the form where the deviation occurs, such as stem, prefix, derivational suffix or inflectional ending
- Orthography: including word boundaries, capitalization, punctuation
- Morphemics: the form includes a non-existing morpheme or a morpheme incompatible with other morphemes present in the form, problems in palatalization, epenthesis or other processes
- Morphology: improper use of a morphological category or word class, also due to agreement or government

In the practical task of manual annotation, it is often difficult to decide what the cause of a specific deviation is. If this is the case, there are two possible strategies: (i) to specify the deviation as occurring at a level where the analysis requires a more sophisticated judgment, i.e. morphosyntax in preference to orthography; or (ii) to specify the deviation in parallel on all relevant levels. We opt for the latter solution, which leaves the decision open for additional analysis and fits well in the concept of cross-classification. In any case, the choice is alleviated by the option of automatic identification of some error types, given a corrected ("emended") form, or even by using a tool suggesting corrections. Actually, the automatic identification always produces at least a formal identification, such as missing or redundant character.

In addition to the layered annotation of the original form, an ill-formed word is assigned a target hypothesis (corrected form) and its analysis, corresponding to the annotation of the original form. Additional categories, such as syntactic function, can also be specified. The two annotation poles – one for the ill-formed and one for the corrected word – are seen as a pattern, a type of mismatch between the annotation layers and the two poles. For a simple case such as (3), the pattern is shown in Table 2.[8] 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 3 represents all cases where a nominative form is used instead of an accusative form.

[7] As Dickinson and Ragheb (2015) also say "There are two main wrinkles to separating linguistic annotation from error annotation, however: 1) annotation categories could employ a notion of grammatical correctness to define; and 2) the decision process for ambiguous cases could reference a sentence's correctness."

[8] In a fully specified pattern, morphological analysis includes all relevant categories, including lemma.

	original	target
formal	replacement of a single character	
location	inflectional suffix	
orthography	*a*	*u*
morphology	nominative noun	accusative noun

Table 2: A pattern for *táta* in (3) (*Lucka viděla *táta*)

	original	target
location	inflectional suffix	
morphology	nominative	accusative

Table 3: An abstract pattern for a form which is nominative instead of accusative

A different type of error is shown in (6). Unlike *táta* in (3), *babičkem* is a non-word. 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.[9]

(6) Byl jsem doma s ***babičkem** → babičkou.
was AUX at home with granny
'I was at home with Grannie.'

The pattern is shown in 4. A more abstract pattern could include only the location and morphemics rows.

	original	target
formal	replacement of two characters	
location	inflectional suffix	
orthography	*em*	*ou*
morphemics	feminine stem + masculine suffix	feminine stem + feminine suffix
morphology	instrumental noun (?)	instrumental noun

Table 4: A pattern for *babičkem* in (6)

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.

4 Identifying non-standard language in a corpus

Except for individual word forms (colloquial, dialectal or non-words) in mainstream corpora and error annotation in learner corpora, corpus annotation rarely distinguishes regular, expected or "standard" expressions on the one hand from less predictable evidence of language use on the other.

Non-standard usage defies general rules of grammar: non-standard language may involve performance errors, creative coinages, or emerging phenomena. Most of these phenomena are still not random, even though it is far from trivial to discover the underlying patterns. 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

[9]The suffix may also be interpreted in other ways, e.g. as the first person plural ending in the present tense of some verbal paradigms (*nesem*). However, rather than multiplying alternatives, which do not appear as likely candidates given the context, we give the author the benefit of the doubt and choose the instrumental interpretation. For the same reason, we refrain from suggesting the hypothesis that the author was at home with her grandpa (*s dědečkem*) rather than her granny (*s babičkou*).

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[10] and checked by a rule-based grammar and valency lexicon, which are also used to infer additional linguistic information about the annotated data.[11] After some detail about the grammar and the lexicon, their current coverage will be presented in terms of statistical results, based on the share of expressions that satisfy the definition of "correctness" as stated by the grammar and the lexicon.

The grammar is used as a watchdog: to check stochastic parses for both formal and linguistic correctness and consistency. Compliant parses receive additional information: relevant properties of lexical categories are projected to phrasal nodes and lexical categories including lemmas matching lexical entries receive valency frames to be saturated by complements in the parse. The grammar is thus supposed to define standard, 'canonical' language in the 'technical' sense of Hirschmann et al. (2007) (see § 2 above). However, this is an idealized picture: the grammar both overgenerates, leaving some non-standard utterances undetected, and undergenerates, falsely deciding that some utterances are not correct – see below for more on this topic.

The grammar consists of a lexical module, providing candidate lexical entries to supply valency frames for verbal lexemes in the data, and a syntactic module, checking the parse and making it more informative by projecting information in the leaf nodes of the constituency tree along the phrasal projections and to the complement sister nodes (dependents). The lexical module operates on lexical entries derived from external valency lexica. The module is responsible for generating a list of entries specific to available diatheses of verbal lexemes. The syntactic module matches the generated lexical entries with the data. The categorial information about words and phrases in the data and in the lexicon follow the cross-classifying taxonomy, used for the learner corpus. The taxonomy captures all distinctions present in a standard Czech tagset used in the stochastic parse and opens the option to use the multi-level scheme to represent non-standard forms in a way it is used in the learner corpus.[12]

The grammar is implemented in Trale,[13] a formalism designed for grammars based on HPSG,[14] a linguistic theory modeling linguistic expressions as typed feature structures. Formally, the grammar consists of two parts: (i) *signature*, i.e. a definition of types of linguistic objects, ordered in a type hierarchy, including their attributes and values; and (ii) *theory*, i.e. a set of constraints on the types and their properties. The parses and lexical entries are in a format compatible with the grammar formalism. The fewer constraints a *constraint-based* grammar includes, the more it *overgenerates*, i.e. the more permissive it is. This is viewed as a welcome property in the development of a grammar that is used primarily for checking existing parses.

There are several important points in which the grammar differs from a standard HPSG grammar, or – more specifically – from a grammar implemented in Trale:

- Rather than parsing or generating strings of word forms, the grammar operates on structures produced by a stochastic parser. As a result, it does not include any syntactic rules of the context-free type. The syntactic backbone, often assumed to be a necessary component of a context-free grammar, is present in the data rather than in the grammar.

- The grammar is run in the mode of a constraint solver, rather than a parser or generator. The constraints come from three sources: the data, the lexicon, and the grammar proper.

- The data are unambiguous in the sense of including a single parse for each sentence. Ambiguities or (preferably) 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.

[10] See Jelínek (2016).

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

[12] See also Petkevič et al. (2015b) for a description of the annotation of periphrastic verb forms using an additional analytical dimension.

[13] `http://www.ale.cs.toronto.edu/docs/`

[14] See, e.g., Pollard and Sag (1994) or Levine and Meurers (2006).

The lexical module uses two external sources of lexical knowledge, both are available and downloadable valency lexicons: VALLEX[15] and PDT-VALLEX,[16] including about 5,000 and 10,000 verbs, respectively, 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) and are used to check whether valency requirements of the verbs in the parsed sentence are met. The lexical module provides the mapping of the frames to their instantiations in specific verbal diatheses and morphological forms, using the same signature and formalism as the syntactic component.

The syntactic module is responsible for checking the parse using the lexical specifications and constraints of the module. The grammar may decide that the parse complies in all respects and provide all available information in the resulting structure. If, however, not all relevant lexical entries are found for the sentence, predicates without valency frames cannot check completeness and coherence of the argument structure in the data, but they can still check grammatical agreement. A valency frame may also cause failure of the check. If so, the sentence is checked also without that frame. A sentence may also fail due to the constraints of the syntactic module. The last remaining and the weakest test is then to apply only the data format definition without constraints (the signature, i.e. the definition of objects and their properties as feature structures representing constituents and their parts).

In most of the above levels of checking, a failure can occur 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 diagnostics 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, the so-called mal-rules.[17]

The examples below illustrate the role of the grammar. In (7-a) and (7-b) the possessive form agrees in gender and case (and number) with the head noun. Examples (7-c) and (7-d) are different: in (7-c) the possessive form does not agree with the head noun in case, in (7-d) in case and gender. Note that the possessive form in (7-c), which is the same as in (7-a), 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 (7-d) has a similar status, but it is acceptable only to speakers of a dialect of Czech.

(7) a. Přitiskl se k otcově noze.
clung RFLX to father's$_{\text{FEM,DAT}}$ leg$_{\text{FEM,DAT}}$
'He pressed against his father's leg.'

b. Přistoupil k otcovu stolu.
approached to father's$_{\text{MASC,DAT}}$ desk$_{\text{MASC,DAT}}$
'He approached his father's desk.'

c. Přistoupil k **?otcově** stolu.
approached to father's$_{\text{MASC,LOC}}$ desk$_{\text{MASC,DAT}}$

d. Přistoupil k **?otcovo** stolu.
approached to father's$_{\text{NEUT,NOM/ACC}}$ desk$_{\text{MASC,DAT}}$

The stochastic parser ignores the agreement mismatch and builds a correct tree. On the other hand, the grammar does not accept the parse. Like every rule-based grammar, the grammar does not have an optimal coverage. In our case it is not a fatal flaw: in most cases a missing account of a phenomenon only means that the grammar is more permissive than it should be (i.e. it overgenerates). The filling of gaps in the coverage is another priority for the future.

The syntactic module includes constraints found in other HPSG-like grammars, such as Head Fea-

[15] See `http://ufal.mff.cuni.cz/vallex`, Lopatková et al. (2008; Žabokrtský and Lopatková (2007)

[16] See Hajič et al. (2003)

[17] 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).

ture Principle, making sure that the head daughter shares appropriate features with its mother, Valency Principle, matching complements and adjuncts with a surface valency frame produced by the lexical component, and other more specific constraints targeting individual types of constructions. The constraints operate mostly on words, such as those specifying morphological categories relevant for agreement within the valency slots of subjects (for subject-predicate agreement) or within a slots for the head (for attribute-noun agreement). The rest is the task of Valency Principle. A special set of constraints is concerned with analytic verb forms, which 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.

A grammar of linguistic competence can never fit the corpus as the evidence of linguistic performance completely. In fact, this may be seen as a benefit: the unexpected or non-standard phenomena in the data can be detected in this way. To distinguish the cases of truly non-standard language from the problems of the syntactic and lexical specifications of the grammar component (useful for the grammar development) on the one hand and to identify and diagnose the types of nonstandard language on the other, the diagnostic part of the tool will be extended to find which specific constraints are violated by which specific words or constructions in the data.

The grammar and lexicon has been developed and tested on a set of 876 sentences, extracted from the manual for the annotation of the Prague Dependency Treebank (Hajič et al., 1997), representing a wide range of linguistic phenomena. Currently, 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. A more extensive testing is under way, using a 100M corpus.

For more extensive testing, the SYN2015 corpus was used, including about 100 million words, i.e. 7.2 million sentences. For 77% sentences at least one valency frame was found and 55% 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 will be to categorize the failures and build a corpus showing the results, including the grammar flags, in a user-friendly way.

5 Discussion and conclusion

We have shown two ways how to approach non-standard languages, with a stress on its proper detection and diagnosis. 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, this is very much work in progress, but we hope to have achieved some promising results already.

Acknowledgements

This work has been supported by the Grant Agency of the Czech Republic, grant no. 16-10185S. The author is grateful to Hana Skoumalová and Jiří Znamenáček for their generous help with the data and to anonymous reviewers for their very useful comments.

References

Katsiaryna Aharodnik, 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*, pages 1428–1436.

Emily M. Bender, 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.

Yevgeni Berzak, 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.

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

Viggo Brøndal. 1928. *Ordklasserne. Partes Orationis.* G. E. C. Gad, København.

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

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

Pitt Corder. 1981. *Error Analysis and Interlanguage*. Oxford University Press, Oxford.

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

Rose-Marie Anne Dechaine. 1993. *Predicates across categories: Towards a category-neutral syntax*. Ph.D. thesis, University of Massachusetts, Amherst.

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

Jakub Dotlačil. 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, September.

Ana Díaz-Negrillo, 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.

Jacob Eisenstein. 2013. What to do about bad language on the internet. In *Proceedings of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 359–369, Stroudsburg, Pennsylvania. Association for Computational Linguistics.

Dan Flickinger 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*, pages 68–73, Sofia, Bulgaria. Association for Computational Linguistics.

Renata Grzegorczykowa, Roman Laskowski, and Henryk Wróbel, editors. 1998. *Gramatyka współczesnego języka polskiego – Morfologia*, volume 1. Wydawniczwo Naukowe PWN.

Jan Hajič, Jarmila Panevová, Eva Buránová, 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. in Czech.

Jan Hajič, 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*, pages 57–68. Växjö University Press.

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

Ray S. Jackendoff. 1977. *X-bar syntax: A study of phrase structure.* MIT Press, Camebridge, Massachusetts.

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

Miroslav Komárek. 1999. Autosemantic Parts of Speech in Czech. In *Travaux du Cercle linguistique de Prague*, volume 3, pages 195–210. Benjamins, Amsterdam.

Steven G. Lapointe. 1999. Dual lexical categories vs. phrasal conversion in the analysis of gerund phrases. In Paul de Lacy and Anita Nowak, editors, *University of Massachusetts Occasional Papers in Linguistics*, number 24, page 157–189. University of Massachusetts.

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

Markéta Lopatková, Zdeněk Žabokrtský, and Václava Kettnerová. 2008. *Valenční slovník českých sloves*. Univerzita Karlova v Praze, Nakladatelství Karolinum, Praha.

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

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

Vladimír Petkevič, 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, editors, *The 5th Workshop on Balto-Slavic Natural Language Processing (BSNLP 2015)*, pages 9–16, Hissar, Bulgaria.

Slav Petrov and Ryan McDonald. 2012. Overview of the 2012 shared task on parsing the web. In *Notes of the First Workshop on Syntactic Analysis of Non-Canonical Language (SANCL)*, volume 59.

Barbara Plank, 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 conjuncion with NAACL 2015)*, pages 148–151. Association for Computational Linguistics.

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

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

Loganathan Ramasamy, 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*, pages 73–80, Prague. Charles University in Prague.

Alexandr Rosen, 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, March.

Alexandr Rosen. 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*, pages 575–590, Olomouc. Palacký University.

Zygmunt Saloni and Marek Świdziński. 1985. *Składnia współczesnego języka polskiego*. Państwowe Wydawnictwo Naukowe, Warszawa.

David Schneider 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, pages 1198–1204, Stroudsburg, PA, USA. Association for Computational Linguistics.

Larry Selinker. 1983. Interlanguage. In *Second Language Learning: Contrastive analysis, error analysis, and related aspects*, pages 173–196. The University of Michigan Press, Ann Arbor, MI.

Petr Sgall and Jiří Hronek. 1992. *Čeština bez příkras*. H&H, Praha.

Mark Steedman and Jason Baldridge. 2011. Combinatory Categorial Grammar. In Robert Borsley and Kersti Börjars, editors, *Non-Transformational Syntax*, pages 181–224. Blackwell.

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