Automatic rule-based correction of stochastic syntactic[1] annotation of Czech

Tomáš Jelínek, Charles University, Prague

1. Introduction

Morphological tagging and lemmatisation have become a standard instrument for exploring large textual corpora. Syntactic annotation, on the other hand, is often limited to small, mainly manually annotated corpora. The aim of our project *Syntactic annotation of Czech corpora* (cf. grant GAČR No. P406/10/0434) is to provide reliable syntactic annotation of large corpora of Czech in which the user will be able to choose the representation of structures and the type of annotation he prefers. In this paper I present a system of automatic annotation, currently under development, which will be used in the project. I demonstrate that even in automatic natural language processing we cannot dispense with a rigorous formal description of the language system. I discuss existing syntactic annotation tools, propose linguistic solutions to recurrent errors in parsing – automatic rule-based correction of syntactic annotation – and I also present preliminary results of tests of this annotation tool.

2. A treebank for all tastes

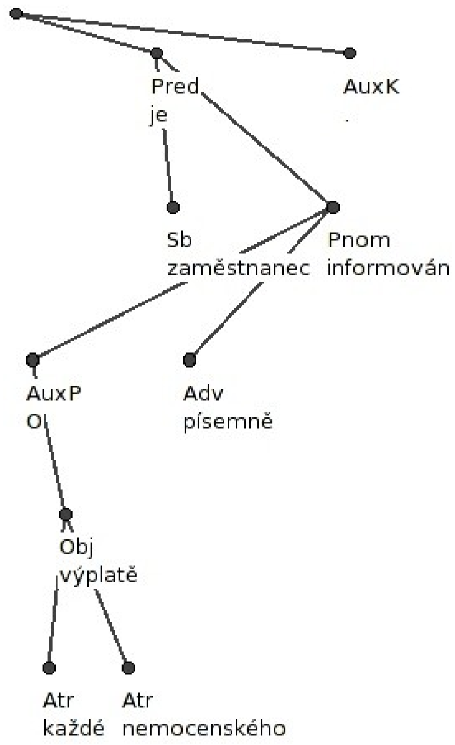
Unlike morphological annotation or lemmatisation, syntactic annotation is much more dependent on theories it is based on, but the interpretation of most of the basic properties of sentences and relations between words are common to many theories, the differences are mostly apparent only in the representation of these structures. The project *Syntactic annotation of Czech corpora* proposes an annotation scheme and a user interface that would enable users to specify their preferences concerning the syntactic representation (e.g. in terms of dependency or constituency) and explore large syntactically annotated corpora, represented according to their preferences. This annotation scheme is described elsewhere in this volume (cf. Jäger et al. 2011), this paper describes the methods of automatic syntactic annotation used in the project.

3. Automatic syntactic annotation

The corpora we wish to annotate syntactically are far too large to be annotated manually (together they contain more than a billion words), so we must resort to automatic parsing. We have neither time nor resources to start from scratch, so we must use existing methods and adapt them to the needs of the project. The best and largest syntactically annotated corpus of Czech as of today is the Prague Dependency Treebank (PDT, cf. Hajič 2006). It contains some 1.5 million manually annotated words (on the syntactic level), has a well-defined formalism that is easily convertible to our internal format and provides an extensive set of tools for automatic tagging and parsing.

[1] This paper was funded by the grant GAČR No. P406/10/0434.

The PDT formalism is based on the Czech linguistic tradition starting with Prague Linguistic Circle, and it is very detailed to encompass all the structures encountered in the 1,5 million corpus. The “analytical layer” of PDT uses a dependency structure with “traditional” syntactic functions: predicate (Pred), subject (Sb), nominal predicate (Pnom), object (Obj), adverbial (Adv), attribute – both agreeing and non-agreeing (Atr) and several auxiliary functions: preposition (AuxP), auxiliary verb (AuxV), coordination (Coord) etc. An example of the graphic representation of the structure is displayed below:



(1) *O každé výplatě nemocenského je zaměstnanec písemně informován*.

‘The employee is informed about each payment of sick leave in writing.’

3.1 Rule-based and stochastic automatic annotation

There are basically two ways in which an automatic annotation can be achieved: stochastic annotation that uses a “training” corpus and probabilities of various structures, and rule-based annotation using formalized linguistic knowledge. For several years, our team has been developing a rule-based morphological disambiguation system and has achieved satisfactory results, but our method of morphological annotation cannot be easily transposed to syntactic annotation: our disambiguation system treats the text for the most part in a “negative” way, eliminating ungrammatical combinations of tags (and lemmas) and leaving only grammatical ones. This negative approach to syntax would require previous generation of all possible structures in the sentence, which would then be trimmed down. However, judging by our experience with morphological disambiguation, it could take years of work to develop the number of rules necessary to obtain exactly one parent node for each node with a better accuracy than the best stochastic parsers.

A positive approach to rule-based parsing of Czech has already been attempted (cf.

Holan & Žabokrtský 2006), but statistical parsers have produced better results. So we decided to use the best stochastic parser available and improve its results using our large database of linguistic data about Czech.

4. Stochastic dependency parser and an analysis of its errors

The single statistical parser for Czech with the best accuracy implemented in the TectoMT system is currently Ryan McDonald's MST Parser with specific features for Czech, cf. McDonald et al. 2005, Novák & Žabokrtský 2007.[2] A detailed description of its function falls outside the scope of this paper, but brief explanation is necessary to understand the proposed system of rule-based corrections.

The key to reliable stochastic annotation is an appropriate setting of the features of the parser, i.e. which linguistic variables (part of speech, case, syntactic function, position in the sentence, lemma) and in what relationship are relevant for the given language. The parser then monitors these variables in the training data, a manually annotated corpus (a part of the PDT treebank), gathering information on the relative frequency of various structures. The result of this training is a large database of probabilities of various partial structures. Once trained, the parser can be used to analyse any morphologically annotated text. It assigns a probability to each possible partial structure and chooses the interpretation with the highest product of partial probabilities. A balance must be sought in the setting of the features, because if the setting is too complex, the amount of information is unmanageable and the parser becomes less reliable.

4.1 The errors of the parser and their causes

To evaluate the performance of the parser, I used it to annotate two large corpora of contemporary Czech, SYN2005 and SYN2010 (both contain approximately 100 million tokens). I manually analysed a sample of 5000 sentences from SYN2005 and ran several automatic tests on both of the corpora to determine whether there was any pattern in the occurrence of errors.

The errors of the parser seem to have two main causes: most of the errors are probably due to the fact that in the annotation phase, the parser cannot encompass the whole sentence, only partial structures (verb – preposition – noun, verb – noun), and is not restricted by its other choices. Therefore the system can calculate a high probability for an ungrammatical structure (two uncoordinated subjects dependent on one finite verb). The second cause is the relatively small volume of the training data: many structures and the majority of less frequent words of the language system never appear in the "training data"; when the parser encounters them in a new text, it does not "know" how to handle them. Some of the errors are caused by erroneous morphological tagging: wrong assignment of case of a noun can lead to wrong assignment of its syntactic function.

[2] See http://ufal.mff.cuni.cz/czech-parsing/, Labeled Accuracy, PDT 2.0. Some parser combinations have achieved slightly better results.

(2) *Ale na přemýšlení jí nezbýval čas.*
wrong: čas$_{\text{accusative/Obj}}$; correct: čas$_{\text{nominative/Sb}}$
‘But she didn’t have time to think.’

4.2 Prepositional phrase as subject: example of a complicated structure

A complex structure too difficult for automatic parsing is appropriately exemplified by prepositional phrases with their head noun labeled as subject. In Czech, these structures are possible, expressing indefinite quantification:

(3)a. *Přes padesát neziskových organizací žádá poslance, aby zákon změnili.*
přes$_{\text{AuxP}}$ *padesát*$_{\text{Sb}}$ *organizací*$_{\text{Atr}}$
‘Over fifty organizations petition the MPs to change this law.’

b. *Ze šesti zemí vzešlo po jednom vítězi.*
po$_{\text{AuxP}}$ *jednom*$_{\text{Atr}}$ *vítězi*$_{\text{Sb}}$
‘One winner came from each of six countries.’

Only a limited set of prepositions can appear in these structures: “na” with accusative, “po” with locative, “přes”, “kolem”, “okolo”, “k”. The verb agrees with a default “neutral” subject: singular (3a) and (3b), neuter (3b). These structures are relatively rare, the parser encounters them only about 50 times in a million words of training data, and the combination of conditions that must be met to allow for such constructions is not used for other syntactic functions (quantification, choice of preposition, verb in neut. sg., etc.). No wonder that the parser makes numerous mistakes in the annotation of similar structures, such as:

(4) *Na zákony, které by uplácení omezily, se stále čeká.*
wrong: *na*$_{\text{AuxP}}$ *zákony*$_{\text{Sb}}$; *correct*: *na*$_{\text{AuxP}}$ *zákony*$_{\text{Obj}}$
‘The laws that would prevent corruption are still to come.’

Moreover, the training data include an even more complicated structure that further confuses the parser:

(5) *Na třicet ázerbájdžánských vojáků a dva Arménci byli zabiti v sobotu během bojů na severovýchodě Náhorního Karabachu.*

na$_{\text{AuxP}}$ *třicet*$_{\text{Sb}}$ *vojáků*$_{\text{Atr}}$ *a dva*$_{\text{Atr}}$ *Arménci*$_{\text{Sb}}$: coordination of a PP subject and a nominal subject

‘Up to thirty Azerbaijani soldiers and two Armenians were killed on Saturday during fighting in the north of Nagorno-Karabakh.’

The verb in example (5) agrees with the second part of the coordination, i.e. masculine animate plural, the parser regards as possible those structures where the verb with a subject in a prepositional phrase is not neuter and singular. We can find such (incorrect) structures in the corpus annotated by the parser:

(6) *Na ostrovy se přeplavili z Jutska, Angelnu a z Dolního Saska.*
wrong: *na*$_{\text{AuxP}}$ *ostrovy*$_{\text{Sb}}$; wrong: *na*$_{\text{AuxP}}$ *ostrovy*$_{\text{Adv}}$
'They crossed over to the islands from Jutland, Angeln and Lower Saxony.'

In example (6), *ostrovy* cannot be interpreted as subject, because there is no quantification in the PP and the verb is in masculine animate plural (and the PP is not coordinated with a noun which would account for this agreement). This kind of detailed analysis of various linguistic phenomena is impossible for the stochastic parser, but necessary for achieving a reliable syntactic annotation.

4.3 Basic syntactic rules regularly violated by the parser

I chose the example of subject PPs for linguistic interest, but the parser commits many more errors in much less complex structures. Several basic syntactic rules for Czech, obvious to a linguist, are regularly violated by the parser. I will present a more sophisticated analysis and more linguistically relevant examples later, but the scope of the paper does not make it possible to discuss all of the rules in detail, so I will enumerate some of them without any further explanation. All of these rules (and many others) can constitute a basis for an automatic correction. For all of these rules, many examples where the parser has violated the rules could be supplied.

Rules

No conjunction can coordinate both syntactic nouns and finite verbs in the same time.
No conjunction can coordinate nouns having different cases.
No syntactic noun or adjective can depend on another sentence constituent across one sentence boundary – the dependency must either remain within the clause, or reach across two or more boundaries (in case of embedded clauses).
Some syntactic functions (subject, nominal predicate, some valency objects) can occur only once in a clause (one finite verb can govern only one subject; one copula can govern a single nominal predicate; most transitive verbs can have only one object in accusative, except *učit* 'teach').
Most verbs can either be reflexive with the pronoun *se*, or have an object in accusative (except *učit* 'teach' again and *dozvědět 'learn, get to know'*).
...

Some of these rules have already been implemented as correction rules, others will be implemented as soon as possible (10 rules have been implemented and tested, other 10 have been implemented, others will follow).

4.4 Automatic identification of errors of the parser

A manual analysis of ten samples of a thousand sentences each annotated by the MST parser has shown that several types of errors are often repeated in the annotation, so I have developed a program to retrieve them automatically in the whole corpus

SYN2005 to identify frequent errors that need to be corrected as a priority. The retrieval algorithm was not very sophisticated, some of the structures identified were actually correct, but it provided a basic outline of the parsing errors. I have been able to identify approximately 4% of the tokens in the corpus as probably erroneous. As the error rate of the parser is between 16% and 20%, about one fifth of the errors of the parser were identified. The following table shows the relative frequency of various errors committed by the parser.

Automatically retrieved errors of McDonald's MST Parser in the corpus SYN2005

Error type	
Verb labeled as main verb in the sentence dependent on another constituent	25.3%
Incompatible syntactic functions coordinated (Adv + Atr, Obj + Atr ...)	13.2%
Two uncoordinated subjects dependent on one verb	11.5%
Pronoun *se* mislabelled as reflexive tantum particle dependent on a transitive verb	10.7%
Noun in nominative labeled as Obj	5.8%
Wrong syntactic function (Adv / Obj) in a PP dependent on a verb (wrong valency)	5.6%
Noun not in nominative labeled as Sb (except for the genitive of negation)	4.2%
Nominal attribute (Atr) dependent on a verb	3.1%
Syntactic nouns in incompatible cases coordinated	1.8%
Noun in accusative labeled as Obj dependent on a non-transitive verb	1.7%
Noun in accusative labeled as Obj dependent on a reflexive verb (except for "učit")	1.4%
Noun in accusative labeled as Obj dependent on a modal/phasal verb with an infinitival object	1.0%

Frequently occurring errors were further manually analysed (samples of 100 erroneous sentences, larger samples if necessary), for most of them a correction algorithm can be developed. It is, however, always much more difficult to correct a wrong structure than merely identify it. Linguistic correction of the results of the stochastic parser is the subject of the next part of this paper.

5. Automatic correction of the results of stochastic parser based on linguistic rules

For many of the frequently recurring error types of the MST Parser, a reliable correction algorithm (a rule) can be found. The parser does not "analyse" the sentence as a linguist would, it lacks the information on clause boundaries, verb valencies etc., it only has a list of relative probabilities of partial structures. It would be possible to include some linguistic information (e.g. valency) directly into the parser features, but it would only make the decision process more complex and not necessarily more successful.

On the other hand, it is possible to develop an independent correction system based on linguistic rules that will operate in a safe way (the system prefers not to correct a possibly erroneous structure if it is too risky). Such a system is flexible, it can be corrected if necessary and extended whenever a new type of errors is found. The system can work together with any parser trained on PDT 2.0, not just the parser I have analysed (although the benefit may be smaller as some types of errors specific for the new parser will not be corrected).

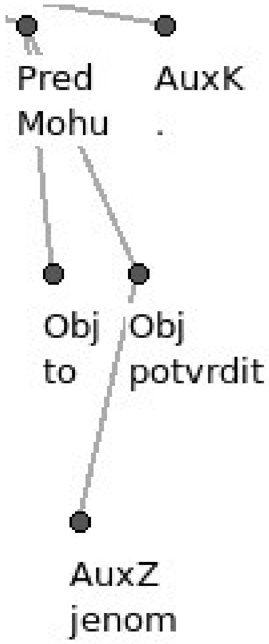
The correction rules have two parts: error identification and error correction. Some types of errors have several correction possibilities depending on the context, sentence boundaries, morphological tags etc. For example, at present erroneous structures with a verb governing two uncoordinated subjects has seven implemented correction options which change the dependency, morphological tag or syntactic function in the sentence, as in (7):

(7) *Debaty, které se už týdny vedou kolem projektu nové Národní knihovny, ...*
original: *které*$_{\text{Sb}}$, *týdny*$_{\text{nominative/Sb}}$; corrected: *které*$_{\text{Sb}}$, *týdny*$_{\text{accusative/Adv}}$
'The debates that have been around for weeks about the project of the new National Library, ...'

In the following three examples of rules implemented in my correction system, I want to show how the correction rules operate and how they use linguistic knowledge. For every rule described, a success evaluation is attached (the correction rules have been applied to the corpus SYN2005 syntactically annotated with the MST Parser, for each correction rule a sample of 200 applications has been examined). Some of the implemented rules have a much more complicated decision algorithm, but there is not enough space to describe them in this paper.

5.1 Object depending on a modal or phase verb

A recurrent type of parser error are structures where a syntactic noun is governed by a modal verb (*moci* 'can', *muset* 'must') or a phase verb (*začít* 'start', *přestat* 'stop') with an infinitival object (if the structure was correct, it would be governed by this infinitive), as in examples (8a,b). For the example (8a), a graphic representation is provided, too:



(8) a. *Mohu to jenom potvrdit.*
pronoun *to*$_{Obj}$ governed by modal *mohu* 'can' instead of *potvrdit* 'confirm'
'I can only confirm that.'
b. *Pravděpodobně o něm budete chtít všem povědět.*
PP o něm$_{Obj}$ governed by *chtít* 'want' instead of *povědět* 'tell'
'You will probably want to tell him all.'

In the case (8b) the correction is straightforward: if the verb in infinitive is a valency verb with the appropriate preposition and case (here *povědět o + locative*), the prepositional phrase is

reattached to the valency verb.
In the case (8a) with an accusative object, a possibility of a wrong tagging must be considered. If following conditions are met, the rule will change the morphological tag and syntactic function and not the dependency on the modal verb::
a) the verb in infinitive is intransitive or it already has another object in accusative
b) it does not have any subject
c) the form of the syntactic noun governed by the modal or phasal verb is ambiguous (accusative – nominative).

(9) *Mohou tyto zkušenosti člověka naplnit?*
original: *zkušenosti*$_{\text{accusative/Obj}}$; corrected: *zkušenosti*$_{\text{nominative/Sb}}$
'Can these experiences fulfil a man?'

5.1.1 Success rate of partial algorithms in the correction rule for an object depending on a modal or phasal verb

subtypes	correction	count	+	0	–
accusative Obj, infinitive is transitive	dependency change	93	100%	0%	0%
PP, infinitive has the right valency	dependency change	30	100%	0%	0%
ambiguous acc/nom.	N4->N1, Obj->Sb	23	76%	23%	0%
TOTAL		146	91%	9%	0%

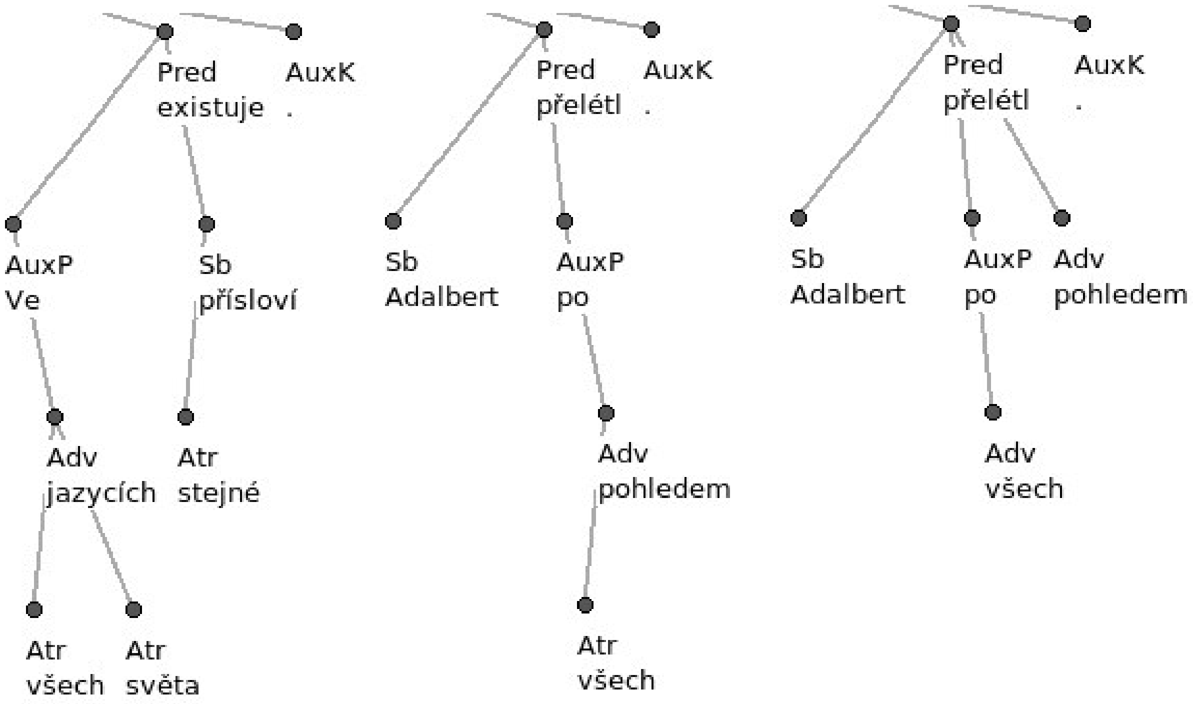
count = number of interventions per 1 million words
+ = percentage of correct interventions that fix the original error completely
0 = percentage of interventions that identify an erroneous structure, change it but fail to correct it in the right way (wrong correction of an error)
– = percentage of wrong interventions that mistake a correct structure for wrong and change it, with a wrong resulting structure

5.2 Incorrect dependency in prepositional phrases with a pronoun

An example of a more complex error type is represented by problems with dependency in structures involving a preposition followed by a pronoun and a noun. In Czech, some pronouns function always as syntactic nouns (*sebe* 'himself', *jemuž* 'who', *nám* 'us'), some function always or mostly as syntactic adjectives (*svůj* 'his', *její* 'her', *nějaký* 'some') and some pronouns function as both, depending on the context (*všem* 'all', *který* 'which, who', *tomu* 'it, that'). In a position directly following a preposition and in front of a noun, a pronoun can
a) depend – as a syntactic noun – on the preposition
b) depend – as a syntactic adjective – on the following noun.
The parser cannot correctly distinguish between pronouns in these categories (the morphological annotation is not much helpful here, as some pronouns with the same tag belong to different categories), and therefore it often handles these structures incorrectly, as in (9b).

Three graphic representations of structures follow, all of them containing the pronoun *všech* ‘all’ in the locative case. In (9a), the structure is correct, the pronoun agrees in number, gender and case with the following noun and does indeed function as an agreeing attribute. In (9b), the structure is incorrect, the pronoun does not agree with the following noun, it functions as an independent syntactic noun, it should be governed by the preposition rather than by the following noun. In (9c) we see the result of the automatic correction of (9b).

(9) a. b. c.



(9) a. *Ve všech jazycích světa existuje stejné přísloví.*
‘In all the languages of the world, the same proverb exists.’

b, c. *Adalbert přelétl po všech pohledem.*
‘Adalbert swept over all of them with his eyes.’

The correction rule checks all structures where a preposition is followed by a pronoun that is either always a syntactic noun, or can possibly be one. In the first case, if the pronoun is not governed by a preposition, the correction rule changes the dependency, so that the pronoun is governed by the preposition and the following noun is governed by the sentence constituent governing the preposition. In the second case, the structure is changed only if the pronoun does not agree in number, gender and case with the following noun, as in (9b, c).

During the evaluation of the rule (see the results below) I found out that the second part of the rule has to be revised: sometimes the correction rule intervened in correct structures with a slightly wrong morphological annotation – the pronoun did not agree in gender with the following noun, even if it formally could. Now the rule takes into account this possibility and corrects the tag instead of the structure.

5.2.1 Success rate of the correction rule for incorrect dependency in PPs

	correction	count	+	0	–
wrong dependency Prep-Pron-Noun	dependency change	77	79%	16%	5%

5.3 Incorrect syntactic function in a prepositional phrase

The last example of error correction is an incorrect assignment of syntactic function of prepositional phrases governed by a verb. With few exceptions it can be either object (when the choice of the preposition and the case depends on the valency of the verb) or adverbial (the preposition and case contribute to the meaning of the adverbial: temporal, local etc.). The boundary is not always clear, but in most cases, it is easy to make the distinction.

As the parser does not include any valency dictionary, it can correctly assign syntactic functions only to prepositional phrases governed by verbs it has encountered in the training data. For the rest, it can only "guess" the most likely function: in Czech, for example the preposition phrases with the preposition *o* and the locative case are mostly objects, with the preposition *v* and the locative mostly adverbials, so the safest course is to always to assign this syntactic function. But in Czech, the syntactic functions of prepositional phrases are not strictly determined by preposition and case: a combination such as *o* and the locative case, typically object, has quite a few examples of adverbials (*o Vánocích* 'during Chrismas') and a combination like *v* and the locative, typically adverbial, is a valency of many verbs (*pokračovat v + loc*, 'continue').

In the correction rules, I use two lists for each preposition and case: a list of valency verbs and a list of typical adverbial prepositional phrases. The correction can go in both ways: *Obj* can be changed to *Adv* or *Adv* changed to *Obj*.

If a PP labeled as *Obj* is a member of the list of typical adverbial PPs and it is governed by a verb that is not a valency verb, the function is changed to *Adv*. If the verb is a valency verb, no change is made.

If a PP is labeled as *Adv,* it is not a member of typical adverbial PPs and it is governed by a valency verb, the syntactic function is changed to *Obj*.

In all the following examples (10), the preposition *v* 'in' with the locative case is used. Twice with a valency verb (*pokračovat* 'to continue'), twice in a typical adverbial collocation (*v roce* 'in the year', *v bezvědomí* 'unconscious'). In the first and in the third example, the parser assigned a wrong syntactic function to the prepositional phrase which was rectified by the correction system, the second example is correct.

(10) a. *Přesto pokračovali v tažení.*
original: *tažení*$_{\text{Adv}}$; corrected*: tažení*$_{\text{Obj}}$ (*v tažení* is not a typical adverbial)
'Nevertheless they continued the campaign.'

b. *V roce 1952 pokračovala likvidace živností opět pomalým tempem...*
‘In the year 1952 the liquidation of small businesses continued at a slower pace.’
correct*: roce*$_{\text{Adv}}$ (*v roce* is a typical adverbial)

c. *Zůstal ležet v bezvědomí.*
original: *bezvědomí*$_{\text{Obj}}$; corrected*: bezvědomí*$_{\text{Adv}}$ (*v bezvědomí* is a typical adverbial)
‘He was lying unconscious.’

My analysis of the results of this correction rule showed that here wrong morphological tagging may occur, too, and must be corrected before any changes of structure. In Czech, several prepositions are used with two (or even three) different cases. If the noun form is ambiguous, too, and the verb governing this prepositional phrase has a valency with the correct preposition but with another case, then if the forms (and possible agreeing attributes) permit it, the case has to be changed:

11. *Většina Bachovy tvorby musela na vydání tiskem čekat víc než století.*
original: *vydání*$_{\text{locative/Obj}}$; corrected*: vydání*$_{\text{accusative/Obj}}$ (*čekat* ‘wait’ is a valency verb with *na + accusative*)
‘Most of Bach’s creation had to wait for the press release more than a century.’

5.3.1 Success rate of partial algorithms in the correction rule of syntactic functions in prepositional phrases

subtypes	correction	count	+	0	–
Adv governed by a valency verb	*Adv->Obj*	1822	96%	0%	4%
Obj governed by a non-valency verb	*Obj->Adv*	878	77%	15%	8%
TOTAL		2700	90%	5%	5%

5.4 Other implemented correction rules and the overall success rate

Seven other correction rules have been already implemented, some twenty others are in preparation. I will briefly describe the function of all the rules already used and tested, then present all the results achieved in one table. The rules are presented in the order of frequency of interventions in the test corpus (SYN2010).

1) The most used correction rule performs only a minor correction of syntactic function of the reflexive pronoun *se*. This pronoun has several functions in Czech: it can be part of a verb reflexive only (*smát se* ‘laugh’), it can be a reflexive object of a transitive verb (*mýt se* ‘wash himself’) or part of the reflexive passive (*auta se vyrábějí* ‘the cars are produced’). Based on a list of reflexive only verbs, this rule checks and corrects the function of the pronoun *se*, if necessary.
2) The rule has been described in detail in 5.3, it checks the syntactic function (Obj or

Adv) of prepositional phrases governed by a verb.

3) The rule corrects structures with two uncoordinated subjects governed by a single verb. The rule either changes the function of one of these subjects (to Pnom; or Obj / Adv changing the tag as well) or the dependency.

4) Another frequent error is a more general problem of dependency: if any syntactic noun is governed by a verb across one or more clause boundaries and it has a finite verb closer (without any sentence boundary between them), if possible, the correction rule changes the dependency of the syntactic noun to the nearest finite verb.

5) The correction rule targets accusative objects governed by a non-transitive verb. It either changes the syntactic function of the object, possibly also its morphological tag, or the dependency.

6) One correction rule focuses on compound prepositions: in PDT, some usual compound prepositions (*na základě* 'based on', *v souvislosti s* 'in connection with') are labelled as such, but the parser does not always identify them, and on the other hand it sometimes labels other prepositional phrases as compound prepositions. The correction rule finds and rectifies these errors.

7) The rule seeks and corrects objects governed by a modal or phase verb, as described in 5.1.

8) The correction rule identifies subordinate clauses labelled as main clauses in the sentence. It changes their function and finds an appropriate governing node.

9) The rule finds and corrects prepositional phrases with a pronoun functioning as a syntactic noun with a wrong dependency, as described in 5.2.

10) The less used implemented rule finds and corrects prepositional phrases incorrectly labelled as subjects (examples in 4.2).

5.4.1 Success rate of all the implemented correction rules

implemented correction rules	correction	count	+	0	–
wrong function of reflexive *se*	AuxT -> Obj / AuxP	5174	64%	26%	10%
wrong syntactic function in PP	Obj->Adv; Adv->Obj	2700	90%	5%	5%
2 x *Sb* governed by one verb	various	995	84%	15%	1%
synt. noun governed by a distant verb	dependency change	794	91%	9%	0%
accusative *Obj* governed by a non-transitive verb	various, e.g. Obj->Sb,N4->N1	290	82%	18%	0%
errors in compound prepositions	dependency change or change of function	203	98%	2%	0%
Obj governed by a modal or phasal verb	various; dependency change	146	91%	9%	0%
subordinate clause labeled as main clause	dependency change	111	89%	7%	4%
wrong dependency in PP with a pronoun	dependency change	77	79%	16%	5%

implemented correction rules	correction	count	+	0	–
wrong syntanctic function (*Sb*) in PP	Sb -> Obj/Adv	17	100 %	0%	0%
TOTAL		10507	85%	12%	3%

count = number of interventions per 1 million words

+ = percentage of correct interventions that fix the original error completely

0 = percentage of interventions that identify an erroneous structure, change it but fail to correct it in the right way (wrong correction of an error)

– = percentage of wrong interventions that mistake a correct structure for wrong and change it, with a wrong resulting structure

5.4.2 Preliminary results of the correction system

After several months of development, the automatic rule-based correction system lowers the error rate of the MST Parser by about 7%. Some wrong interventions occur, but they can be relatively easily corrected by expanding the dictionaries used by the rules or by improving the correction algorithms. The system is open, new rules can and will be inserted when reliable correction algorithms are found.

6. Conclusion

In our project, we aspire to create a large treebank with a reliable annotation and a friendly interface that will allow users to choose their preferred representation of syntactic structures. In order to achieve a reliable syntactic annotation, we need a good annotation method. The MST Parser I described in this paper provides good results, but they can be considerably improved by linguistic knowledge applied. Correction rules based on linguistic properties unavailable to the parser (valency, clause boundaries etc.), using extensive lists of words with specific properties can rectify many shortcomings of the stochastic annotation. Only 10 implemented rules decreased the error rate in the analysed samples of the annotated corpus by 7%.[3] Ten others have already been implemented (without any precise analysis of the results), many others will be added in the course of the project. We hope that the new *Treebank for all tastes* will be a useful research tool.

References

Czech national corpus – SYN2005. 2005. *SYN2010. 2010.* Praha: Ústav Českého národního korpusu FF UK. Available from WWW: <http://www.korpus.cz>.

Hajič, Jan et al. 2006. *Prague Dependency Treebank 2.0.* CD-ROM, Philadelphia: Linguistic Data Consortium.

Holan, Tomáš, Zdeněk Žabokrtský. 2006. Combining Czech Dependency Parsers. In *Lecture Notes In Computer Science: Proceedings of the 9th International Conference, TSD.* Berlin, Heidelberg: Springer-Verlag, 95-102.

[3] The improvement in labeled accuracy (correct parent, correct syntactic function) is an estimate based on tests of the performance of the correction rules and the frequency of their application on the corpus SYN2005. The output of the correction rules is not yet fully compatible with the test data of the PDT, so no independent testing was possible.

Jäger, Petr, Vladimír Petkevič, Alexandr Rosen & Hana Skoumalová. 2011. Towards a treebank for all tastes. In *this volume*.

Kocek, Jan, Marie Kopřivová & Karel Kučera, eds. 2000. *Český národní korpus – úvod a příručka uživatele*. Praha: FF UK – ÚČNK.

McDonald, Ryan, Fernando Pereira, Kiril Ribarov & Jan Hajič. 2005. Non-projective dependency parsing using spanning tree algorithms. In *HLT'05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing.* Vancouver. 523–530.

Novák, Václav & Zdeněk Žabokrtský. 2007. Feature Engineering in Maximum Spanning Tree Dependency Parser. In *Proceedings of the 10th International Conference on Text, Speech and Dialogue*. *Plzeň: Západočeská univerzita*. Berlin, Heidelberg: Springer-Verlag, LNCS 4629, 92–98.