Post-annotation Checking of the Treebank

Barbora Hladká, Petr Pajas

April 30, 2001

This document contains a list of tests and subsequent corrections in the data (the morphological and syntactic analytic layer annotations) of the Prague Dependency Treebank 1.0 obtained from the annotators. The first two sections discuss the corrections done separately on each individual layer. The corrections based on a mutual revision of the morphological vs. syntactic analytic annotations are described in the last section.

1 Morphological Layer

In the framework of the post-annotation checking of the morphologically annotated data, for each word token, we compare manually assigned (*lemma, MTag*[1]) pair with the output of the automatic morphological analyzer (AMA). To classify the annotation as a correct one with regard to the current AMA, the manual (*lemma, MTag*) pair must correspond to exactly one (*lemma, MTag*) pair out of all possible pairs determined by the AMA.

For example, in a particular context, word token *stav* is annotated as a noun (`NNIS4-----A----`) with lemma *stav*; by comparison of both manual information and the AMA output listed in Fig. 1, the annotation exactly fits in the AMA output.

```
<f>stav<l>stav<t>NNIS4-----A----
<MMl>stav<MMt>NNIS1-----A----<MMt>NNIS4-----A----
<MMl>stavět<MMt>Vi-S---2--A----<MMt>Vi-S---3--A---4A
<MMl>stavit_:T_:W<MMt>Vi-S---2--A----<MMt>Vi-S---3--A---4
<MMl>stát-2_:W_^(něco_se_přihodilo)<MMt>VmYS------A----
```

Figure 1: Annotation vs. AMA

We are aware that chosen checking strategy gives us only the incorrect *word form* annotations (so-called *visible* ones) - those annotations which do not have their 'counterpart' in the AMA output (see Fig. 2 - `<f>stav<l>stav<t>NNIS6-----A----`); those annotations which are at the same time *incorrect* with regard to the context and *correct* with regard to the possible morphological readings of a particular word form remain hidden (see Fig. 3 - `<f>stav<l>stav<t>NNIS1-----A----`).

The annotation of word forms unknown to the AMA serves as a way of improving (with regard to the number of recognized word forms) of the AMA.

Besides the visible errors mentioned above, we had to revise the word forms with a manual tag "`X@-------------`" (i.e. such word forms have never been recognized by the AMA). All discovered misspelled strings were replaced by the correct word forms (determined by the context information) and the original incorrect word forms were stored in the SGML element *w* via the

[1] morphological tag

```
<f>popiš<l>popsat_:W<t>Vi-S---2--A----
<MMl>popsat_:W<MMt>Vi-S---2--A----<MMt>Vi-S---3--A---4
```

```
<f>stav<l>stav<t>NNIS6-----A----
<MMl>stav<MMt>NNIS1-----A----<MMt>NNIS4-----A----
<MMl>stavět<MMt>Vi-S---2--A----<MMt>Vi-S---3--A---4A
<MMl>stavit_:T_:W<MMt>Vi-S---2--A----<MMt>Vi-S---3--A---4
<MMl>stát-2_:W_^(něco_se_přihodilo)<MMt>VmYS------A----
```

Figure 2: Incorrect *Word Form* Annotation

```
<f>popiš<l>popsat_:W<t>Vi-S---2--A----
<MMl>popsat_:W<MMt>Vi-S---2--A----<MMt>Vi-S---3--A---4
```

```
<f>stav<l>stav<t>NNIS1-----A----
<MMl>stav<MMt>NNIS1-----A----<MMt>NNIS4-----A----
<MMl>stavět<MMt>Vi-S---2--A----<MMt>Vi-S---3--A---4A
<MMl>stavit_:T_:W<MMt>Vi-S---2--A----<MMt>Vi-S---3--A---4
<MMl>stát-2_:W_^(něco_se_přihodilo)<MMt>VmYS------A----
```

Figure 3: Incorrect *Context* Annotation

attribute *spell* (<`w spell`>). The discovered missing words have been added and such a new word token is preceded by the SGML mark-up <`w ins`>. Similarly, the discovered words which were by mistake divided into more than one word as well as the words which were (also by mistake) joined into a single word were replaced by the proper words and the original mistakes are stored in the SGML element *w* via the attribute *phrpart* and *ctcd*, respectively.

Altogether, within the post-annotation checking of the morphological annotations, we have passed a sequence of the following checking steps **twice**:

1. processing of the morphologically annotated data by the AMA
2. manual evaluation of the annotations vs. the AMA output visible discrepancies
3. the AMA improving

2 Syntactic Analytic Layer

The list of corrections on the syntactic-analytic layer covers only linguistics-related topics when many technical-like ones that affected neither the tree structure nor syntactical tags are ommited. The tests were intended to help us locate the most evident mistakes that the annotators, authors or programs could have made during the the process of annotation. Note the fact that a sentence fails certain test *does not* by itself mean it is wrong or misannotated.

There are also many other tests (not included in this list) that were not yet used but should be used in order to make the date consistently annotated in many ways according to the Annotator's Manual.

The order of the items does not correspond with the order in which the tests were actually applied.

List of the post-annotation tests:

1. Annotators' messages written in a special attribute of data nodes were considered and appropriate corrections were made where needed.

2. [AuxK] is placed on the very end of a sentence. Fail of this test usually means that there are more sentences within a single tree. This is caused by a mistake in the automated process of dividing text into sentences. There is a dual problem of one sentence divided into several trees. None of these problems may be automatically corrected (or searched). However since we corrected it each time we bumped upon it while searching other problems, there should not be many instances of this left.

3. The misspelled strings or words which were by mistake divided into more than one node (word) as well as words which were (also by mistake) joined into a single node (word) are highlighted during the checking of the morphological annotation. Corrections of such words/nodes were made. This test is also one of the ways used to search for nonsensical or non-annotable parts of data (like huge tables of numbers, blocks of graphical symbols such as rules, TV programs etc.) which were pruned in reasonable cases.

4. [pnom] depends on 'být' ('to be')

5. [Obj] seldom depends on 'být'

6. Nodes depending on the root (#) may take one and only one tag of this set: [Pred], [Coord], [Apos], [AuxC], [AuxK], [AuxG], [Exd], [AuxP], while [AuxP] is allowed only in the [ExD]-constructions. Moreover, there is only one [Pred] depending on [AuxK] directly and several other similar conditions must hold.

7. There are no dependent nodes of [AuxV], [AuxG], [AuxO], [AuxK], [AuxR], [AuxT], [AuxX]

8. There are rare cases in which there is a node dependent on [AuxY]. The only permitted tag for such a node is [AuxY]

9. [*_Co] must be a part of some [Coord] construction (in most cases its son, although there may be [AuxP] and/or [AuxC] constructions in between)

10. Same as (9) but applied on [*_Ap] and [Apos]

11. Only [AuxZ] may depend on [AuxZ]

12. [Pred] is always son of root (#) unless there is [AuxC] between them

13. Nouns tagged with [Sb]$^{*)}$ are (almost) always in 1st case

14. Words 'a','však' are either [Coord]$^{*)}$ or [AuxY]

15. Words 'avšak','nebo' are [Coord]$^{*)}$

16. Comma (,) is [AuxX], [Coord]$^{*)}$ or [Apos]$^{*)}$

17. [AtvV] depends on a verb

18. [Atv] does not depend on a verb

19. [AuxS] belongs only to roots (#) of the trees

20. Special (unspecified) tags [???] were replaced by correct values.

21. Nodes or trees tags `[---]` intended for non-annotable parts of text were pruned in reasonable cases.

22. Comma ',' is `[AuxX]`, `[Coord]` or `[AuxY]`*) and it has sons if it is not `[AuxX]`

23. Words 'což' (usually `[Sb]` or `[Obj]`) and 'přičemž' (`[Adv]`) depend on a coordinated predicate

24. There must be at least one node with the `_Co` or `_Ap` suffix in afun under each coordinating node (with afun Coord or Apos resp.). Between the coordinating node and its descendant with the suffix may only be nodes with afuns AuxC and AuxP, i.e. prepositions and conjunctions.

25. Some particular functions (e.g. `[AuxO]`, `[AuxS]`, `[AuxT]`, `[AuxV]`, `[AuxZ]`, `[Coord]`, `[AuxX]`, `[AuxP]`, `[AuxC]`, `[AuxK]`, `[AuxG]`*)) can be assigned only to members of defined sets of words.

26. Infinitives under modal verbs are `[Obj]` (or `[Sb]` under 'lze'). Auxiliary future forms of 'být' are under infinitives of main verbs.

27. There can not be more than one `[Sb]` under a verb.

28. 'zatímco' is `[AuxC]`, 'přitom' is `[Adv]`

29. 'jakoby' is `[AuxY]` if under a verb, `[AuxZ]` otherwise

30. if 'být' is not `[AuxV]`, there is usually something dependent on it

31. 'však' and 'ale' are always `[Coord]`*)

32. 'proto' is never `[AuxC]`

33. 'ano', 'ne', 'ba' are never `[AuxZ]`

34. 'stále' and 'jestě' are sisters nad `[Adv]` (in collocation 'stàle ještě')

35. 'jako je', 'jako jsou' etc. in appositional meaning have 'jako' dependent as `[AuxY]` on the form of 'být' marked as `[Apos]`

36. All components of graphical expressions as '1)', 'A .' etc. (meaning 'At first etc.') are sisters.

37. No larger sequence of sentences appears more than once in the corpora

38. 'jako' as `[AuxY]` is never a sister of an `[Atv]`

39. 'nemluvě' is `[Adv_Pa]` (in collocation 'nemluvě o')

40. 'soudě' is `[Adv_Pa]` (in collocation 'soudě podle')

41. 'raději', 'radši' are `[Pnom]` under 'být', `[Adv]` or `[Atv]` elsewhere

42. 'počínaje', 'konče' are usually `[AuxP]` if they precede their dependent node, otherwise they are `[Exd_Pa]` or seldom `[Atv]`

43. 'neřkuli' is always `[AuxZ]`

44. no 'form' attribute longer than one character can end with '.'

45. in collocation 'jako by(chom/sme/ste/ch/s)', 'jako' is `[AuxC]` and 'by(...)' is `[AuxV]`

46. ‘dokud’ is always [Adv], while ‘pokud’ can be [AuxC] as well

47. in collocation ‘přesto, že’, ‘přesto’ and ‘že’ are sisters, where ‘přesto’ is [Adv] and ‘že’ is [AuxC]

48. ‘než’ is either [AuxC] or [ExD]

49. in collocation ‘přeci/přece jen/jenom’, the first word depends on the latter and both are [AuxY]

50. Title and the first sentence of the first paragraph are not glued together (can be detected only in some cases, e.g. the title is all capitalized)

51. in collocation ‘ať už/již’, ‘už/již’ is dependent and [AuxY], ‘ať’ is [AuxC]

3 Morphological Layer vs. Syntactic Analytic Layer

In this step the annotation on the two layers (the morphological and analytic) was compared to a certain degree. The comparison concentrated on the following aspects:

1. The most important nodes in the analytical layer (namely nodes which were annotated by the analytical function (afun) Obj (object), Sb (subject) and Pred (predicate)) were tested against their morphological tag.

   For example:

   - The annotation of subject (afun Sb) passed the test as correct if tagged either as a noun, pronoun, numeral or adjective in Nominative, verb in the infinitive form, or a literal number. For subjects depending on a subordinating conjunction the test failed unless the subject was tagged as a verb form on the morphological layer.
   - The annotation of object (afun Obj) failed the test if tagged as a form in Nominative.
   - Nodes assigned afun Pred were allowed to pass the test if and only if tagged as a verb form.

2. Prepositional phrases: prepositions and secondary prepositions were tested for presence in a list given in the Manual. Also, the case of a preposition was tested against the case of a depending noun, pronoun, numeral or adjective.

3. Agreement in case, gender and number between predicate and depending subject, as well as between attribute and its governing node was checked.

*) the test applies also on analytical tags with one of the suffixes _Ap,_Co,_Pa