J. Hlaváčová (Ed.): ITAT 2017 Proceedings, pp. 181–185
CEUR Workshop Proceedings Vol. 1885, ISSN 1613-0073, © 2017 T. Jelínek

FicTree: a Manually Annotated Treebank of Czech Fiction

Tomáš Jelínek

Charles Univeristy, Faculty of Arts,
Prague, Czech Republic
Tomas.Jelinek@ff.cuni.cz

Abstract: We present a manually annotated treebank of Czech fiction, intended to serve as an addendum to the Prague Dependency Treebank. The treebank has only 166,000 tokens, so it does not serve as a good basis for training of NLP tools, but added to the PDT training data, it can help improve the annotation of texts of fiction. We describe the composition of the corpus, the annotation process including inter-annotator agreement. On the newly created data and the data of the PDT, we performed a number of experiments with parsers (TurboParser, Parsito, MSTParser and MaltParser). We observe that the extension of PDT training data by a part of the new treebank actually does improve the results of the parsing of literary texts. We investigate cases where parsers agree on a different annotation than the manual one.

1 Introduction

The Czech National Corpus (CNC) has decided to enrich the annotation of some of its large synchronous corpora by syntactic annotation, using the formalism of the Prague Dependency Treebank (PDT) [4]. The parsers used for syntactic annotation must be trained on manually annotated data, with only PDT data available now. To achieve a reliable parsing, it is necessary to ensure the training data to be as close as possible to the target text, but in PDT, the texts are only journalistic, while one third of the texts in representative corpora of synchronous written Czech of the CNC belongs to the fiction genre. In many ways, fiction differs considerably from the characteristics of journalistic texts, for example by a significantly lower proportion of nouns versus verbs: in the journalistic genre, 33.8% tokens are nouns and 16.0% are verbs; in fiction, the ratio of nouns and verbs is almost equal, 24.3% tokens are nouns, and 21.2% verbs (based on statistics [1] from the SYN2005 corpus [3]).

Therefore, a new manually annotated treebank of fiction texts was created; it was annotated according to the PDT a-layer guidelines. The scope of the new treebank is only about 11% of the PDT data, due to the difficulties of manual syntactic annotation, but even so, using this new resource does improve the parsing of fiction texts.

In this article we present this new treebank, named FicTree (*Tree*bank of Czech *fict*ion), its composition, and the annotation process. We describe the first experiments with parsers based on the data of FicTree and PDT. In the data of the FicTree treebank parsed by four parsers, we investigate cases where all parsers agree on a syntactic annotation of one token which differs from the manual annotation.

2 Composition of the Treebank

The manually annotated treebank FicTree is composed of eight texts and longer fragments of texts from the genre of fiction published in Czech from 1991 to 2007, with a total of 166,437 tokens, 12,860 sentences. It is annotated according to the PDT a-layer annotation guidelines [5]. The PDT data annotated on the analytical layer comprise, for comparison, 1,503,739 tokens, 87,913 sentences. Seven of the eight texts which compose the FicTree treebank, were included in the CNC corpus SYN2010 [7] (the eigth one was originally intended to be included in the SYN2010 corpus too, but was removed in the balancing process). The size of the eight texts ranges from 4,000 to 32,000 tokens, the average is 20,800 tokens. Most of the texts are written in original Czech (80%), the remaining 20% are translations (from German and Slovak). Most of the texts belong to the fiction genre without any subgenre (according to the classification of the CNC), one large text (18.2% of all tokens) belongs to the subclass of memoirs, 5.9% tokens come from texts for children and youth.

The language data included in the PDT and in FicTree differ in many characteristics in a similar way to the differences between the whole genres of journalism and fiction described above. In FicTree, there are significantly shorter sentences with an average of 12.9 tokens per sentence compared to an average of 17.1 tokens per sentence in PDT. The part-of-speech ratio is also significantly different, as shown in Table 1.

It is evident from the table that there is a significantly lower proportion of nouns, adjectives and numerals in FicTree, and a higher proportion of verbs, pronouns and adverbs, which corresponds to the assumption that in fiction, verbal expressions are preferred, whereas journalism tends to use more nominal expressions.

3 Annotation Procedure

The FicTree treebank was syntactically annotated according to the formalism of the analytical layer of the Prague Dependency Treebank. The texts were lemmatized and morphologically annotated using a hybrid system of rule-based desambiguation [6] and stochastic tagger Featu-

Table 1: POS proportion in PDT and FicTree

	PDT	FicTree
Nouns	35.60	22.31
Adjectives	13.72	7.73
Pronouns	7.68	16.42
Numerals	3.83	1.53
Verbs	14.34	23.16
Adverbs	6.18	9.19
Prepositions	11.39	9.14
Conjunctions	6.61	9.39
Particles	0.64	1.05
Interjections	0.01	0.07
Total	100	100

rama[1]. The texts were then doubly parsed using two parsers: MSTParser [9] and MaltParser [10] (the parsing took place several years ago when better parsers such as TurboParser [8] were not available) trained on the PDT a-layer training data. The difference in the algorithms of both parsers ensured that the errors in the texts were distributed differently, it can be assumed that errors in the subsequent manual corrections will not be identical. According to Berzak [2] there are likely some deviations common for both parsers, which will also manifest in the final (manual) annotation, but this distortion of the data could not be avoided.

3.1 Manual Correction of Parsing Results

The automatically annotated data was then distributed to three annotators that checked one sentence after using the TrEd software for manual treebank editing and corrected the data. The two versions of the parsed text (parsed by the MSTParser and by the MaltParser) were always assigned to two different annotators, we ensured that the combinations of parsers and annotators were varied. The data were divided into 163 text parts of approx. 1000 tokens, every combination of parsers and annotators has occurred in at least 10 text parts (the proportion of texts corrected by indivudual annotators was 26%, 35% and 39%).

The task of the manual annotators was to correct syntactic structure and syntactic labels, but they also had the possibility to suggest corrections of segmentation, tokenization or morphological annotation and lemmatization.

3.2 Adjudication

The two corrected versions of syntactic annotation from each text were merged, the resulting doubly annotated texts were examined by an experienced annotator (adjudicator) who decided which of the proposed annotations to accept. The adjudicator was not limited to the two manually corrected versions, she was allowed to choose another solution consistent with the PDT annotation manual and data. Some changes in tokenization and segmentation were also performed (159 cases, mainly sentence split or merge). The adjudication took approximately five years of work due to the difficulty of the task, the effort to maximize the consistency of the same phenomenon across the treebank (and in accordance with PDT data), and other workload with a higher priority.

3.3 Accuracy of the Parsing and of the Manual Corrections

In the following two tables, we will present the accuracy of each step of annotation and the inter-annotator agreement. Table 2 shows to what extent the automatically parsed and the manually corrected versions of the text agree with the final syntactic annotation, first for the texts annotated with the MSTParser, then for the ones annotated with the MaltParser. Two measures of agreement with the final annotation are shown: UAS (unlabeled attachment score, i. e. the proportion of tokens with a correct head) and LAS (labeled attachment score, i. e. the proportion of tokens with a correct head and dependency label).

Table 2: Accuracy of annotated versions

	UAS:auto.	UAS:man.	LAS:auto.	LAS:man.
MST	83.37	96.92	75.31	95.03
Malt	86.08	96.40	79.39	94.42

It is clear from the table that due to the relatively low input parsing quality, the annotators had to carry out a large number of manual interventions in the parsing correction process. The dependencies or labels were modified for 15–20% of tokens. The manually corrected versions differ much less from the final annotation, the disagreement is approx. 5% of the tokens.

Table 3 presents the agreement between the two automatically parsed versions and the inter-annotator agreement (the agreement between the two manually corrected versions). As in the previous table, we use the measures UAS and LAS.

Table 3: Agreement between parsers and inter-annotator agreement

	UAS	LAS
Parsers	83.48	75.66
Annotators	93.89	90.26

The table shows that the agreement between the automatically annotated versions is very similar to the agree-

[1] See http://sourceforge.net/projects/featurama.

ment between the final annotation and the worse of the two parsing results. After the manual corrections, the agreement between the two versions of texts has increased considerably, but the difference is approximately twice the difference between each of the manually corrected versions of texts and final syntactic markings. This fact shows that the final annotation alternately used the solutions from both versions of the texts.

4 Parsing Experiments

We conducted a series of experiments on PDT and FicTree data. All data was automatically lemmatized and morphologically tagged using the MorphoDiTa tagger [12].[2] We used four parsers, two parsers of older generation, which were used for the automatic annotation of FicTree data (before manual corrections, with a different morphological annotation and with other settings providing a better parsing accuracy): MSTParser [9][3] and MaltParser [10];[4] and two newer parsers: TurboParser [8][5] and Parsito [11].[6] We use three measures: UAS (unlabeled attachment score), LAS (labeled attachment score) and SENT (labeled attachment score for whole sentences, i. e. the proportion of sentences in which all tokens have correct heads and syntactic labels).

4.1 Training on the PDT Data

The first experiment was to compare the parsing of the PDT test data (journalism) and the whole FicTree data (fiction) using parsers trained on PDT training data (journalism). The results of the experiment are shown in Table 4. Two following columns compare the results on the PDT etest and on the whole FicTree data.

Table 4: Accuracy of parsers trained on PDT train data

	UAS	UAS	LAS	LAS	SENT	SENT
	etest	FicTree	etest	FicTree	etest	FicTree
MST	85.93	84.91	78.85	76.82	23.79	26.94
Malt	86.32	85.01	80.74	77.94	31.32	31.86
Parsito	86.30	84.62	80.78	77.65	31.17	31.32
Turbo	88.27	86.66	81.79	79.06	27.74	29.61

The results of the experiment with the UAS and LAS scores for all parsers are approximately 2% worse for FicTree than for PDT, probably due to the genre differences of FicTree versus PDT data. In the case of SENT, the FicTree scores are comparable or better than the PDT etest, probably because the sentence length in FicTree is significantly lower, so there is a higher percentage of well-parsed sentences.

4.2 Training on PDT Data Combined with FicTree

In the second experiment, we split FicTree data into training data (90%) and test data (10%) and combined the FicTree training data with the PDT training data. This experiment was repeated three times with different distribution of the FicTree data, in order to achieve a more reliable result (10% of FicTree is only 16,000 tokens). In that way, 30% of FicTree has effectively been used as test data, the parsers beeing trained on PDT training data plus each time 90% of FicTree. It would have been better to use the whole FicTree data in a 10-fold cross-validation experiment (always adding 90% of data to train PDT and testing the remaining 10%), but we lacked the time and computational resources to do so. Table 5 compares the results of parsers trained on the PDT training data itself and on these merged data (train+ in the table), using PDT etest data and FicTree test data. For each of the measures (UAS, LAS, SENT), the accuracy of the parser trained on the PDT training data is always in one table column, in the following column, there is the accuracy measured for the parser trained on the combined training data (PDT and FicTree, train+). The average for the three experiments is shown.

Table 5: Accuracy of parsers trained on PDT train data (train) and PDT&FicTree train data (train+)

	UAS	UAS	LAS	LAS	SENT	SENT
Etest	train	train+	train	train+	train	train+
MST	85.93	85.98	78.85	78.90	23.79	23.23
Malt	86.32	86.41	80.74	80.87	31.32	31.62
Parsito	86.30	86.48	80.78	81.02	31.17	31.53
Turbo	88.27	88.34	81.79	81.89	27.74	27.93
	UAS	UAS	LAS	LAS	SENT	SENT
FicTree	train	train+	train	train+	train	train+
MST	85.03	85.49	77.24	77.68	26.78	27.18
Malt	85.10	87.14	78.25	81.39	28.92	36.14
Parsito	84.81	86.42	77.99	80.53	31.01	36.52
Turbo	87.00	88.35	79.69	81.69	29.12	34.92

It is clear from the table that extending the training data by a part of the FicTree treebank is beneficial both for parsing the PDT test data and for parsing FicTree data. The improvement in the parsing of the PDT etest is not statistically significant (approximately 0.05% for UAS), but it

[2] Available on http://ufal.mff.cuni.cz/morphodita.

[3] Available on https://sourceforge.net/projects/mstparser/. Used with the parameters: decode-type:non-proj order:2.

[4] Available on http://www.maltparser.org/. Used with the stacklazy algorithm, libsvm learner and a set of optimized features obtained with MaltOptimizer.

[5] Available on http://www.cs.cmu.edu/~ark/TurboParser/. Used with default options.

[6] Available on https://ufal.mff.cuni.cz/parsito. Used with hidden_layer= 400, sgd= 0.01,0.001, transition_system= link2, transition_oracle= static.

is consistent for all parsers and measures except for the measure SENT for the MSTParser.

For the FicTree test data, we note a significant improvement in parsing, the increase in the measures is between 0.4% and 2.5%. It is therefore clear that for the syntactic annotation of texts of fiction, the extension of the training data by the FicTree training data is definitely beneficial.

5 The Agreement of Parsers versus the Manual Annotation

We also attempted to use the results of the parsing to assess the quality of the manual annotation and adjudication of the FicTree treebank. The whole FicTree data was annotated by four parsers trained on the PDT training data. From these parsed data, we chose those cases where all four parsers agree on one dependency relation and / or syntactic function of a token, whereas the manual syntactic annotation is different. In total, parsers agreed for 70.04% of tokens in the FicTree data (78.12% if we only count dependencies without syntactic labels). 5.17% of all tokens do not match manual annotation (3.43% of tokens without syntactic labels). Table 6 shows 10 syntactic functions which occur most frequently in such cases of agreement between four parsers and disagreement with manual annotation. In the first column, the syntactic label from the manual annotation is shown. In the second column, we present the proportion of disagreement in the tokens with this syntactic label, in the third column, there is the absolute number of occurrences.

Table 6: Syntactic labels where parsers agree with each other but disagree with manual annotation

Synt. label	Ratio	Number
Adv	5.49	1135
Obj	6.20	1065
AuxX	6.08	618
Sb	5.64	561
ExD	13.96	543
AuxC	11.65	536
AuxP	4.05	501
Atr	1.76	339
AuxV	8.08	302
AuxY	15.85	271

The data in the table shows that differences between parsers and manual markup often occur with the Adv and Obj syntactic labels (adverbial and object), since the annotation performed by parsers often differs from the manual annotation due to the difficulty of linguistic phenomena. Frequent differences between parsing results and manual annotations are discussed in more detail later, we will first give two examples of such differences and their supposed reason.

5.1 Examples of Differences between Manual Annotation and Parsing Results

The first example, a sentence fragment *pohledy plné bezměrné důvěry*, 'regards full of unbounded trust' displayed below, shows a typical example of wrong parsing result due to incorrect morphological annotation. The parsers agree on an erroneous interpretation of the syntactic structure. After the tokens where dependencies or syntactic labels differ, we show the annotation (numbers indicate relative differencies, –1 means that the governing node is positioned 1 to the left, +2 governing node is 2 to the right; syntactic labels are shown if they differ).

Pohledy plné/–1/+2 bezměrné důvěry/Obj/-2/Atr/–3
Regards full of unbounded trust

Incorrect morphological tagging of the ambiguous form *plné* 'full' (which can formally agree both with the preceding noun *pohledy* 'regards' and with the following noun *důvěry* 'trust' in number, gender and case) led the parsers to ignore the valency characteristics of the adjective *plný* 'full', they consider it to be the attribute of the following noun *důvěry* 'trust', which they interpret as a nominal attribute of the preceding noun *pohledy* 'regards'. The manual annotation is correct, the adjective *plný* 'full' is dependent on the preceding noun *pohledy* 'regards', the following noun *důvěry* 'trust' is an object of the adjective. Similar differences in the attribution of the Adv and Obj syntactic labels and their dependency relations are common, the manual annotation is in most cases correct (the parsers agree on an erroneous syntactic structure).
In some cases, it is unclear whether the manual annotation or the parsing results are correct, as in the following sentence:

Doktorka/+6/+1 vychutnávala chvíli efekt svých slov a pak pokračovala:
The doctor enjoyed for a while the effect of her words, and then went on:

The head of the subject *Doktorka* 'doctor' in manual annotation is the coordinating conjunction *a* 'and' which coordinates two verbs representing two clauses: *vychutnávala* 'enjoyed' and *pokračovala* 'went on/continued'. The subject is considered as a sentence member modifying the whole coordination (i. e. both verbs). However, all parsers agree on a different head: the verb *vychutnávala* 'enjoyed' closest to the subject. In this interpretation, the second verb has a null subject (pro-drop). Both interpretations are possible in the formalism of PDT, there is no strict rule indicating when the subject should modify coordinated verbs and when it should depend on the closest verb only. In the PDT data, both solutions are used. (The more the structures in the coordinated sentences are similar and simple, the more likely it is that the subject will be common.).

5.2 Most Frequent Discrepancies between Parsing Results and Manual Annotation

In cases where dependencies between the manually assigned one and the one on which the parsers agree are different, the syntactic labels are usually the same. These functions are mostly auxiliary functions: AuxV (auxiliary verbs), AuxP (prepositions) and AuxC (conjunctions) or are related to punctuation (AuxX, AuxK, AuxG). When the syntactic labels differ, the most frequent mismatches are Obj and Adv, Sb and Obj, Adv and Atr.

The highest proportion of discrepancies between the manually and automatically assigned functions is related to the following functions: AuxO (46.5%), AuxR (21.9%), AuxY (15.9%), ExD (14.0%) and Atv (13.5%). AuxO and AuxR refer to two possible syntactic functions of the reflexive particles *se/si* 'myself, yourself, herself...' depending on context, for correct parsing, understanding of semantics and use of lexicon would be necessary. The AuxY function covers particles and other auxiliary functions, ExD is a function which covers several different phenomena in the PDT formalism and is difficult to parse automatically. None of these functions occur frequently in the training data.

5.3 Manual Analysis

When we analyzed manually a sample of sentences in which four parsers agree on a dependency or syntactic label different from the one chosen manually, we found out that in 75% of cases, the manual annotation was certainly correct, about 20% of the occurrencies could not be decided quickly due to the complexity of the construction, in less than 5% of such occurrences the manual annotation was incorrect. It would certainly be useful to carefully check all cases of such discrepancy, it may reduce the error rate in FicTree data by about 0.2–0.5%, but for now we lack the resources to do so.

6 Conclusion

The new manually annotated treebank of Czech fiction FicTree will allow for a better syntactic annotation of texts of fiction when we add it to the PDT training data. Given that larger training data were shown to be beneficial in parsing journalistic texts as well, its use may be broader. We plan to publish the FicTree trebank in the Lindat / CLARIN repository in the near future (after additional checks of selected phenomena) and we would like to publish it later in the Universal Dependencies[7] format, too, using publicly available conversion and verification tools.

Acknowledgement

This paper, the creation of the data and the experiments on which the paper is based have been supported by the Ministry of Education of the Czech Republic, through the project Czech National Corpus, no. LM2015044.

References

[1] T. Bartoň, V. Cvrček, F. Čermák, T. Jelínek, V. Petkevič: "Statistiky češtiny /Statistics of Czech". NLN, Prague, 2009.

[2] Y. Berzak, Y. Huang, A. Barbu, A. Korhonen, B. Katz: "Bias and Agreement in Syntactic Annotations", in Computing Research Repository, 1605.04481, 2016.

[3] F. Čermák, D. Doležalová-Spoustová, J. Hlaváčová, M. Hnátková, T. Jelínek, J. Kocek, M. Kopřivová, M. Křen, R. Novotná, V. Petkevič, V. Schmiedtová, H. Skoumalová, M. Šulc, Z. Velíšek: "SYN2005: a balanced corpus of written Czech". Institute of the Czech National Corpus, Prague, 2005. Available on-line: http://www.korpus.cz.

[4] J. Hajič: "Complex Corpus Annotation: The Prague Dependency Treebank," in Šimková M. (ed.): Insight into the Slovak and Czech Corpus Linguistics, pp. 54–73. Veda, Bratislava, Slovakia, 2006.

[5] J. Hajič, J. Panevová, E. Buráňová, Z. Urešová, A. Bémová, J. Štepánek, P. Pajas, J. Kárník: "A manual for analytic layer tagging of the prague dependency treebank." ÚFAL Internal Report, Prague, 2001.

[6] T. Jelínek, V. Petkevič: "Systém jazykového značkování současné psané češtiny," in Čermák F. (ed.): Korpusová lingvistika Praha 2011, vol. 3: Gramatika a značkování korpusů, pp. 154-170. NLN, Prague, 2011.

[7] M. Křen, T. Bartoň, V. Cvrček, M. Hnátková, T. Jelínek, J. Kocek, R. Novotná, V. Petkevič, P. Procházka, V. Schmiedtová, H. Skoumalová: "SYN2010: a balanced corpus of written Czech". Institute of the Czech National Corpus, Prague, 2010. Available on-line: http://www.korpus.cz.

[8] A.F.T. Martins, M.B. Almeida, N.A. Smith: "Turning on the Turbo: Fast Third-Order Non-Projective Turbo Parsers," in Proceedings of ACL 2013, 2013.

[9] R. McDonald, F. Pereira, K. Ribarov, J. Hajič: "Non-projective Dependency Parsing using Spanning Tree Algorithms," in Proceedings of EMNLP 2005, 2005.

[10] J. Nivre, J. Hall, J. Nilsson: "MaltParser: A Data-Driven Parser-Generator for Dependency Parsing," in Proceedings of LREC 2006, 2006.

[11] M. Straka, J. Hajič, J. Straková, J. Hajič jr.: "Parsing Universal Dependency Treebanks using Neural Networks and Search-Based Oracle," in Proceedings of TLT 2015, 2015.

[12] J. Straková, M. Straka, J. Hajič: "Open-Source Tools for Morphology, Lemmatization, POS Tagging and Named Entity Recognition," in Proceedings of ACL 2014, 2014.

[7] See universaldependencies.org.