An Annotated Corpus Outside Its Original Context: A Corpus-Based Exercise Book

Barbora Hladká and Ondřej Kučera
Institute of Formal and Applied Linguistics, Charles University
Malostranské nám. 25
118 00 Prague
Czech Republic
hladka@ufal.mff.cuni.cz, ondrej.kucera@centrum.cz

Abstract

We present the STYX system, which is designed as an electronic corpus-based exercise book of Czech morphology and syntax with sentences directly selected from the Prague Dependency Treebank, the largest annotated corpus of the Czech language. The exercise book offers complex sentence processing with respect to both morphological and syntactic phenomena, i. e. the exercises allow students of basic and secondary schools to practice classifying parts of speech and particular morphological categories of words and in the parsing of sentences and classifying the syntactic functions of words. The corpus-based exercise book presents a novel usage of annotated corpora outside their original context.

1 Introduction

Schoolchildren can use a computer to chat with their friends, to play games, to draw, to browse the Internet or to write their own blogs - why should they not use it to parse sentences or to determine the morphological categories of words? We do not expect them to practice grammar as enthusiastically as they do what is mentioned above, but we believe that an electronic exercise book could make the practicing, which they need to do anyway, more fun.

We present the procedure of building an exercise book of the Czech language based on the Prague Dependency Treebank. First (in Section 2) we present the motivation for building an exercise book of Czech morphology and syntax based on an annotated corpus – the Prague Dependency Treebank (PDT). Then we provide a short description of the PDT itself in Section 3. Section 4 is the core of our paper. Section 4.1 is devoted to the filtering of the PDT sentences in such a way that the complexity of sentences included in the exercise book exactly corresponds to the complexity of sentences exercised in traditional Czech textbooks and exercise books. Section 4.2 documents the transformation of the sentences – more precisely a transformation of their annotations into the school analysis scheme as recommended by the official framework of the educational programme for general secondary education (Jeřábek and Tupý, 2005). The evaluation of the system is described in Section 4.3. Section 5 summarizes this paper and plans for the future work.

2 Motivation

From the very beginning, we had an idea of using an annotated corpus outside its original context. We recalled our experience from secondary school, namely from language lessons when we learned morphology and syntax. We did it "with pen and paper" and more or less hated it. Thus we decided to build an electronic exercise book to learn and practice the morphology and the syntax "by moving the mouse around the screen."

In principle, there are two ways to build an exercise book - manually or automatically. A manual procedure requires collecting sentences the authors usually make up and then process with regard to the chosen aspects. This is a very demanding, time-consuming task and therefore the authors manage to collect only tens (possibly hundreds) of sentences that simply cannot fully reflect the real usage of a language. An automatic procedure is possible when an annotated corpus of the language is available. Then the disadvantages of the manual procedure dis-

appear. It is expected that the texts in a corpus are already selected to provide a well-balanced corpus reflecting the real usage of the language, the hard annotation work is also done and the size of such corpus is thousands or tens of thousands of annotated sentences. The task that remains is to transform the annotation scheme used in the corpus into the sentence analysis scheme that is taught in schools. In fact, a procedure based on an annotated corpus that we apply is semi-automatic, since the annotation scheme transformation presents a knowledge-based process designed manually - no machine-learning technique is used.

We browsed the Computer-Assisted Language Learning (CALL) approaches, namely those concentrated under the teaching and language corpora interest group (e.g. (Wichmann and Fligelstone (eds.), 1997), (Tribble, 2001), (Murkherjee, 2004), (Schultze, 2003), (Scott, Tribble, 2006)). We realized that none of them actually employs manually annotated corpora – they use corpora as huge banks of texts without additional linguistic information (i.e. without annotation). Only one project (Keogh et al., 2004) works with an automatically annotated corpus to teach Irish and German morphology.

Reviewing the Czech electronic exercise books available (e.g. (Terasoft, Ltd., 2003)), none of them provides the users with any possibility of analyzing the sentence both morphologically and syntactically. All of them were built manually.

Considering all the facts mentioned above, we find our approach to be novel one. One of the most exciting aspects of corpora is that they may be used to a good advantage both in research and teaching. That is why we wanted to present this system that makes schoolchildren familiar with an academic product. At the same time, this system represents a challenge and an opportunity for academics to popularize a field with a promising future that is devoted to natural language processing.

3 The Prague Dependency Treebank

The Prague Dependency Treebank (PDT) presents the largest annotated corpus of Czech, and its second edition was published in 2006 (PDT 2.0, 2006). The PDT had arisen from the tradition of the successful Prague School of Linguistics. The dependency approach to syntactic analysis with the main role of a verb has been applied. The annotations go from the morphological layer through to the intermediate syntactic-analytical layer to the tectogrammatical layer (the layer of an underlying syntactic structure). The texts have been annotated in the same direction, i. e. from the simplest layer to the most complex. This fact corresponds with the amount of data annotated on each level – 2 million words have been annotated on the lowest morphological layer, 1.5 million words on both the morphological and the syntactic layer, and 0.8 million words on all three layers.

Within the PDT conceptual framework, a sentence is represented as a rooted ordered tree with labeled nodes and edges on both syntactic (Hajičová, Kirschner and Sgall, 1999) and tectogrammatical (Mikulová et al., 2006) layers. Thus we speak about syntactic and tectogrammatical trees, respectively. Representation on the morphological layer (Hana et al., 2005) corresponds to a list of (word token and morphological tag) pairs. Figure 1 illustrates the syntactic and morphological annotation of the sample sentence *Rozdíl do regulované ceny byl hrazen z dotací.* [The variation of the regulated price was made up by grants.] One token of the morphological layer is represented by exactly one node of the tree (*rozdíl* [variation], *do* [of], *regulované* [regulated], *ceny* [price], *byl* [was], *hrazen* [made up], *z* [by], *dotací* [grants], '.') and the dependency relation between two nodes is captured by an edge between them, i. e. between the dependent and its governor. The actual type of the relation is given as a function label of the edge, for example the edge (*rozdíl, hrazen*) is labeled by the function *Sb* (subject) of the node *rozdíl*. Together with a syntactic function, a morphological tag is displayed (*rozdíl*, *NNIS1-----A----*).

Since there is *m:n* correspondence between the number of nodes in syntactic and tectogrammatical trees, it would be rather confusing to display the annotations on those layers all together in one tree. Hence we provide a separate tree visualizing the tectogrammatical annotation of the sample sentence – see Figure 2. A tectogrammatical lemma and a functor are relevant to our task, thus we display them with each node in the tectogrammatical

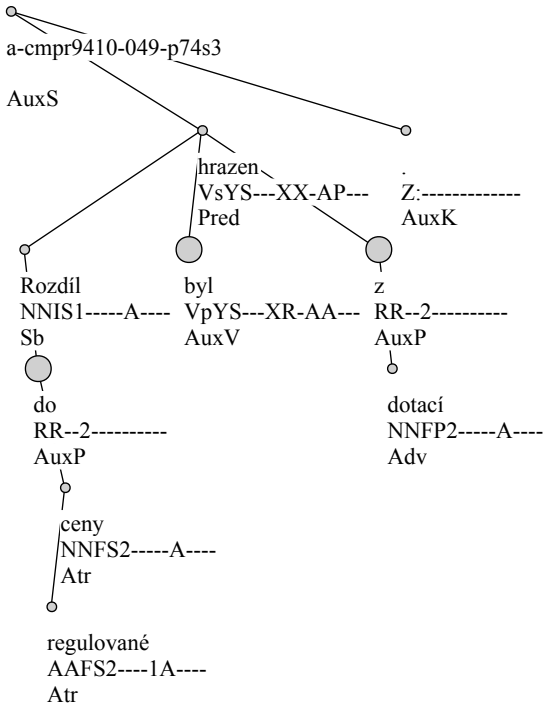

Figure 1: A PDT syntactic tree of the sentence *Rozdíl do regulované ceny byl hrazen z dotací.*

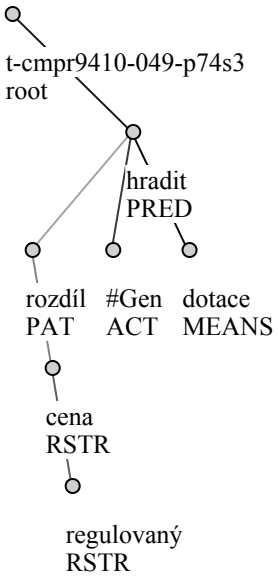

Figure 2: A PDT tectogrammatical tree of the sentence *Rozdíl do regulované ceny byl hrazen z dotací.*

tree, e. g. (*hradit*, *PRED*).

In the following text, we will be using the term the *PDT approach* when having in mind the conceptual framework of PDT annotation, and the *school approach* when having in mind the conceptual framework of a sentence analysis as it is taught in schools.

4 Exercise book composition

With regards to our idea, the electronic exercise book is an electronic system that consists of

- a database of sentences with their morphological and syntactic analyses automatically generated from an annotated corpus,
- a user interface
  - to select sentences from the database or, in other words, to compose the exercises,
  - to simultaneously analyze the selected sentences both morphologically and syntactically,
  - to check the analyses.

More specifically, the composition of the PDT-based exercise book of Czech morphology and syntax implies the selection of those sentences from PDT that are annotated morphologically and syntactically. However, there emerge some syntactic phenomena that are handled differently in the PDT approach than in the school approach. The data annotated tectogrammatically has to be taken into account to process these phenomena properly. Given that, the data annotated on all three layers (0.8 million words in 49,442 sentences) become the *candidate set* of sentences from which the exercise book is to be composed.

Unfortunately, the sentences from the candidate set cannot be merely taken as they are because of two factors:

- the complexity of sentences in the PDT goes

beyond the complexity of sentences in textbooks;

- some syntactic phenomena are handled differently in the PDT approach than in the school approach.

This means that some of the sentences have to be completely discarded (sentence filtering, see 4.1) and syntactic trees of the remaining sentences have to be transformed to match the school analysis of syntax (see 4.2). Luckily, the school approach to the morphology does not coincide with the PDT approach. Therefore the PDT morphological annotations do not need any special handling. It is impossible to browse the candidate set of sentences manually with regard to its volume. Both *sentence filtering* and *annotation transformation* must be done automatically. The whole process is shown in Figure 3.

To summarize, our work on the electronic exercise book covers the data and the software components ((Hladká, Kučera, 2005), (Kučera, 2006), (STYX, 2008)):

- *Annotated Sentence Database* Almost 12,000 annotated sentences generated by the *FilterSentences* component.

- *FilterSentences*. A component used to prepare the annotated sentence database suitable for usage in the exercise book. The end user will never have to use this.

- *Charon*. An administrative tool, used for viewing all of the available sentences and for composing the exercises. We assume that mostly teachers will use it.

- *Styx*. The electronic exercise book itself. It uses the exercises composed with Charon. An active sentence is analyzed both morphologically and syntactically as shown in Figure 4. During the morphological analysis, the user moves word by word, and for each word selects its part of speech. According to the selected part of speech, the combo boxes for the relevant morphological categories appear and let the user choose one of several choices they consider the proper one. During the syntactic analysis, the user moves nodes using the traditional drag and drop method to catch the dependent-governor relation. Afterwards, the syntactic functions are assigned, technically via pop-up menus. Once the analyses are finished, the correct answers are provided separately for morphology and syntax.

4.1 Sentence filtering

The candidate set consists of many sentences that are not appropriate for schoolchildren to analyze. They contain phenomena that authors of textbooks either do not consider at all or sometimes do not agree upon. The following seven filtering criteria have been formulated to exclude problematic sentences. For each filter, we provide a brief description.

1. *SimpleSentences.* The most complex filter that excludes compound and complex sentences.

2. *GraphicalSymbols.* Excludes sentences with various graphical symbols (except for the dot sign) because they imply more complex phenomena than the school analyses operate with.

3. *EllipsisApposition.* Excludes sentences containing an ellipsis or an apposition.

4. *OnePredicate.* Excludes sentences without a predicate (sentences with more than one predicate are already excluded by *SimpleSentences*).

5. *LessThanNWords.* Excludes sentences that are too long.

6. *MoreThanNWords.* Excludes sentences that are too short (usually simple headlines).

7. *AuxO.* Excludes sentences containing emotional, rhythmic particles carrying the *AuxO* syntactic function.

The filters were applied in the same order as they are listed above. First the filter *SimpleSentences* was applied on the candidate set of sentences. Then the sentences preserved by this filter were filtered by *GraphicalSymbols*, and so on. Table 1 provides an overall quantitative overview of sentence filtering – for illustration, the most complex filter *SimpleSentences* excluded the highest percentage of sentences

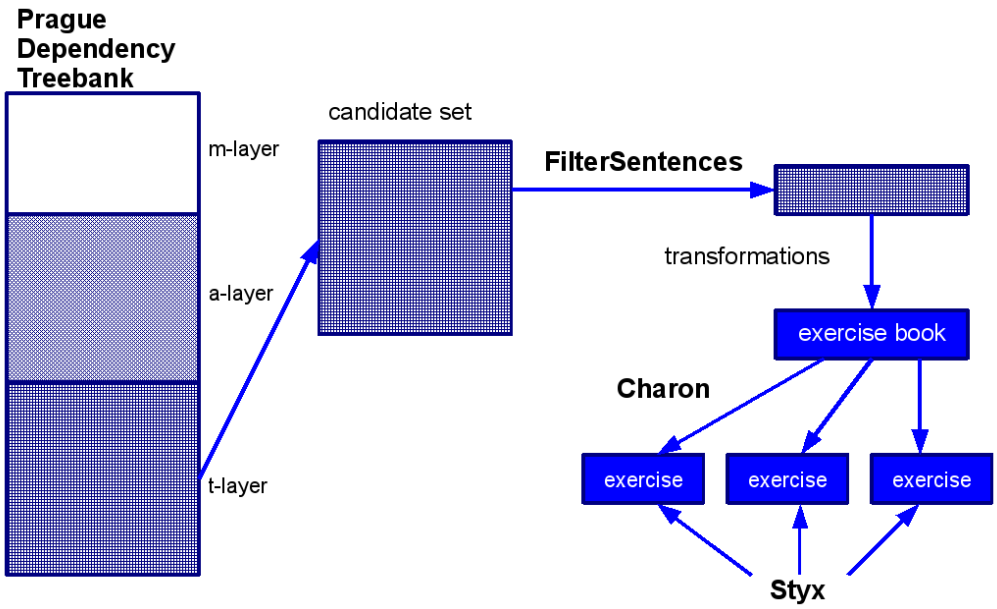

Figure 3: Exercise book composition

(54.4 %). As it is highlighted in the last row of the table, almost 12,000 sentences were preserved after processing the candidate set with all the filters.

Filter	# input sentences	# preserved sentences (%)
SimpleSentence	49,442	22,552 (45.6)
GraphicalSymbols	22,552	20,384 (90.4)
EllipsisAposition	20,383	13,633 (66.9)
OnePredicate	13,633	13,617 (99.9)
LessThanNWords	13,617	13,010 (95.5)
MoreThanNWords	13,010	11,718 (90.1)
AuxO	11,718	11,705 (99.9)
overall	**49,442**	**11,718 (23.7)**

Table 1: Quantitative overview of sentence filtering

4.2 Annotation transformation

In the school approach, a sentence is represented as a tree-like structure with labeled nodes. Unlike PDT syntactic trees, the structures of the school approach have no root node or, in another point of view have two roots: a subject and a predicate (see Figure 5 – *rozdíl, byl hrazen* respectively).

Besides the above-mentioned difference in analysis schemes, the PDT and the school approach differ in the following aspects:

- Many of the PDT syntactic functions do not have counterparts in the school approach.

- The school approach does not have the direct 1:1 correspondence between nodes of the morphological layer and the syntactic layer, i. e. a node can contain more than just one word as visible in Figure 5 – the pair of words *byl, hrazen* form one node as well as the pair *z, dotací*. The words inside each node are listed in accordance to the surface word order of the sentence.

With regards to the discussed differences, we systematically went through the PDT annotation guidelines (Hajičová, Kirschner and Sgall, 1999), analyzed all specified phenomena and designed their transformations into the school analysis scheme. Three elementary operations on syntactic trees and the rules mapping syntactic functions have been formulated. Then a transformation is understood as a sequence of these operations and mapping rules.

1. *JoinTheParentNode* The words at the node are moved up to the parent node and all child nodes of the given node become the child nodes of the parent node. The node is removed afterwards.

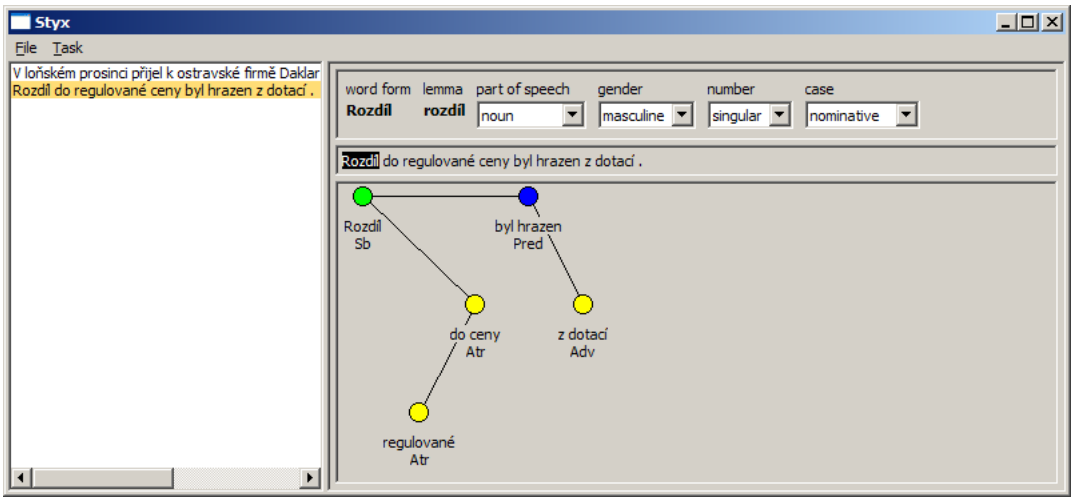

Figure 4: Styx—practicing

2. *AbsorbTheChildNodes* The words at all child nodes of the node are moved into the node. The child nodes are removed and their child nodes become the child nodes of the node. This operation is equivalent to the *JoinTheParentNode* operation applied to all child nodes of the node.

3. *RemoveTheNode* The node-leave is removed.

Mapping PDT syntactic functions follows these operations on trees. Given the complexity of syntactic phenomena and the differences between the approaches, it is not possible to map all functions in a straightforward way as is evident from Table 2. While the school approach works with seven syntactic functions (listed in the second column) the PDT approach labels with 25 functions[1] (listed in the first column). The PDT functions indicating the subject, the predicate, the attribute and the adverbial (in italics) are simply mapped to their school counterparts. The other functions are changed into the school functions in accordance with the type of operation the nodes they belong to pass. After the *AbsorbTheChildNodes* operation, the node is mostly labeled by the direct school counterpart of its "most important child node", i.e. the child node bearing one of the simply-mapped functions, vaguely noted. After the *JoinTheParentNode* operation, the parent node does not change its function in most cases.

PDT syntactic functions	school syntactic functions	description
Pred	Přs	predicate
Pnom	Přj	predicate nominal
Sb	Po	subject
Obj	Pt	object
Atr, AtrAdv, AdvAtr, AtrAtr, AtrObj, ObjAtr	Pk	attribute
Adv, Atv, AtvV	Pu	adverbial
Obj	D	complement
Coord	—	coordination
AuxC, AuxP, AuxZ, AuxO, AuxV, AuxR, AuxY, AuxK, AuxX, AuxG	—	auxiliary sentence members

Table 2: School vs. PDT syntactic functions

For illustration, a PDT syntactic tree in Figure 1 is transformed into a school structure displayed in Figure 5. Needed transformations include, for example. merging the nodes *(do, AuxP)* and *(ceny, Atr)* into the node *(do ceny, Pk)* or similarly merging *(byl, AuxV)* and *(hrazen, Pred)* into *(byl hrazen, Přs)*.

4.3 Evaluation

It is always difficult to evaluate such systems. It is impossible to express the quality of our system with

[1] The total number of the PDT syntactic functions is actually higher. Here we list those functions that appear in sentences included in the exercise book.

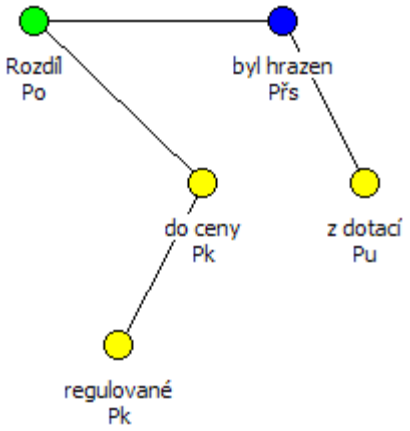

Figure 5: A school syntactic tree of the sentence *Rozdíl do regulované ceny byl hrazen z dotací.*

numerical figures only. The only number we can provide presents the sentence count included in the exercise book: We believe that almost 12,000 sentences bring enormous diversity to the practicing of morphology and syntax.

To find out the real value of our system, we presented it to two different audiences. First we presented it to academics, who really appreciated the idea of corpus assimilation for morphology and syntax learning in schools. Their discussions were mainly concerned with the transformation of annotations.

Then we presented the exercise book during Czech classes in secondary schools. We found out that both the teachers and the students were immediately able to use the system and they were excited about it. They agreed that such exercises would be a nice addition to their classes. Given the experience we acquired during the presentations, we created a sample class (a methodological guide) for teachers, and we collected some interesting ideas that may help us improve the system. These improvements concern i) the annotation transformations (1, 2, 3); ii) the variety of exercises (4); iii) the user interface (5):

1. We do not distinguish between the different types of adverbials. Thus we will provide the possibility of marking a node as being a place adverbial or time adverbial etc.

2. We do not distinguish concordant and discordant attributes yet.

3. Dealing with coordination needs revision, especially when it comes to a difference between dependents of the coordination as a whole and dependents of members of the coordination.

4.
   - During the morphological analysis, the user selects only the part of speech of the given word and STYX itself provides the relevant morphological categories to analyze. In this fashion, the exercises are too simplistic. To master the morphology, the user must know which categories are relevant to the given part of speech.
   - The Charon module will give the user the option of selecting sentences that contain some specific phenomena. Currently, an administrator goes through all the sentences ”manually” and if they fulfill her/his selection criteria, (s)he includes them in the exercises.

5. The user interface has to be changed to be more ”crazy,” or dynamic, to attract not only the ”A” pupils but the rest of them as well. Much more comfortable controls, for example by adding keyboard shortcuts for the most common actions, will be offered too.

5 Conclusion

The PDT-based exercise book has completed its initial steps. The theoretical aspects have been analyzed, the system has been implemented and demonstrated to schoolchildren. Their feedbacks motivates us to improve the system in such a way that it will become a real educational tool.

References

Hana Jiří and Dan Zeman and Hana Hanová and Jan Hajič and Barbora Hladká and Emil Jeřábek. 2005. A Manual for Morphological Annotation, 2nd edition. *ÚFAL Technical Report 27*, Prague, Czech Republic.

Hajičová Eva and Zdeněk Kirschner and Petr Sgall. 1999. A Manual for Analytic Layer Annotation of the Prague Dependency Treebank (English translation). *ÚFAL Technical Report*, Prague, Czech Republic.

Hladká Barbora and Ondřej Kučera. 2005. Prague Dependency Treebank as an exercise book of Czech. In *Proceedings of HLT/EMNLP 2005 Interactive Demonstrations*, pp. 14-15. Vancouver, British Columbia, Canada.

Jeřábek Jaroslav and Jan Tupý 2005. *The official framework educational programme for general secondary education.* Research pedagogical institute, Prague.

Keogh Katrina and Thomas Koller and Monica Ward and Elaine UíDhonnchadha and Josef van Genabith 2004. CL for CALL in the Primary School. In *Proceedings of the International Workshop in Association with COLING 2004*, Geneva, Switzerland.

Kučera Ondřej. 2006. Pražský závislostní korpus jako cvičebnice jazyka českého. *Master thesis*. Charles University, Prague, Czech Republic.

Mikulová Marie et al. 2006. A Manual for Tectogrammatic Layer Annotation of the Prague Dependency Treebank. *ÚFAL Technical Report*, Prague, Czech Republic.

Mukherjee, J. 2004. Bridging the gap between applied corpus linguistics and the reality of English language teaching in Germany. In U. Connor and T. Upton (eds.) *Applied Corpus Linguistics: A Multidimensional Perspective*. Amsterdam: Rodopi, pp. 239-250.

PDT 2.0 [online]. 2006. *Prague Dependency Treebank, 2nd edition.* http://ufal.mff.cuni.cz/pdt2.0

Scott Mike and Christopher Tribble 2006. *Textual Patterns: keyword and corpus analysis in language education*. Amsterdam: Benjamins.

Schultze Mathias 2003. AI in CALL: Artficially Inated or Almost Imminent? In *Proceedings of the World-CALL Conference*, Banff, Canada.

STYX [online]. 2008. *The STYX electronic exercise book of Czech* http://ufal.mff.cuni.cz/styx

Terasoft, Ltd. 2003. *TS Český jazyk 2 - jazykové rozbory*. http://www.terasoft.cz.

Tribble Christopher 2001 Corpora and teaching: adjusting the gaze. In *Proceedings of the ICAME 2001 Conference*, Louvain, Belgium.

Wichmann Anne and Steven Fligelstone (eds.) 1997. *Teaching and Language Corpora (Applied Linguistics and Language* London: Addison Wesley Longman.