Getting more data – Schoolkids as annotators

Jirka Hana, Barbora Hladká

Charles University in Prague
Faculty of Mathematics and Physics
Czech Republic
{hana, hladka}@ufal.mff.cuni.cz

Abstract

We present a new way to get more morphologically and syntactically annotated data. We have developed an annotation editor tailored to school children to involve them in text annotation. Using this editor, they practice morphology and dependency-based syntax in the same way as they normally do at (Czech) schools, without any special training. Their annotation is then automatically transformed into the target annotation schema. The editor is designed to be language independent, however the subsequent transformation is driven by the annotation framework we are heading for. In our case, the object language is Czech and the target annotation scheme corresponds to the Prague Dependency Treebank annotation framework.

Keywords: annotated corpora, morphology, syntax, crowdsourcing

1. Introduction

The classic quote *"the more (annotated) data, the better"* is being examined every time when the supervised learning techniques are applied, regardless the language in questions (e.g., (Brants *et al.*, 2007). Textual data annotation is a task expensive in terms of time, expertise, and money. Thus new alternative ways of annotation are searched for at present. *Games With A Purpose* (von Ahn and Dabbish, 2008) represent the most popular alternative way of annotation. They exploit the capacity of Internet users who like to play on-line games. Moreover, the players work simply by playing the game – the data are generated as a by-product of the game. The game popularity brings more game sessions and thus more annotated data.

The GWAP methodology was formulated in parallel with design and implementation of on-line games with images, which enjoy enormous popularity. So far, a number of GWAP with texts have been designed (e.g., (Hladká *et al.*, 2011) but they do not enjoy popularity as great as games with images mainly because reading a text is less fun than observing images.

Our novel approach to "cheaper" annotation involves the idea of including school children (and teachers, if interested) into the annotation process. We provide them with a simple annotation editor and very basic training (few minutes). Apart from that, we do not teach them the usual annotation guidelines and instead we rely on their knowledge of syntax and morphology as taught in school. Their analyses are transformed into the desired annotation scheme automatically.

Obviously, we do not expect that they will do it so enthusiastically as they play on-line games. On the other hand, they have to (or at least should) practice grammar of language anyway so we can expect some annotation.

2. Parsing sentences at schools

In the Czech Republic, children at the elementary and high-school level are required to parse sentences into dependency trees. They are trained in sentence diagramming that is based on dependency syntax of Šmilauer (Šmilauer, 1972), see Figure 1. Since the Slovak education curriculum has the same roots as the Czech one, Slovak school children undertake a similar training.

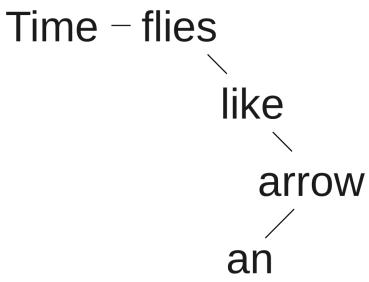

Figure 1: Sample of the Czech diagram

We did research based on personal communication to see whether there are similar requirements on children in other countries, that is, eventually, whether we could broaden our methodology to other languages. We asked mostly our colleagues at universities and we can see that the schoolchildren are not trained in such activity in most countries. If they are, it is usually limited to marking phrases such as subject or object, without identifying their hierarchical structure. The initiative by Richard Hudson (Hudson, 1992) is devoted to the same research.[1]

In some English speaking countries, the so-called Reed-Kellogg sentence diagraming (Kellogg and Reed, 1899) was fairly common in the past. However, it has mostly disappeared from the current curricula (even though there have been some attempts to reintroduce recently). Czech sentence diagrams (trees) look graphically different from Reed-Kellogg's scheme, but formally they are very similar, compare Figure 1 and 2.

[1] http://www.phon.ucl.ac.uk/home/dick/ec/school-grammar.htm

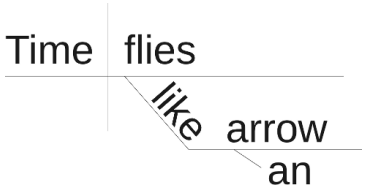

Figure 2: Reed-Kellogg sentence diagramming

3. Computational linguistics goes to Czech schools

We have developed a suite of applications (see Figure 3) aimed at those who want to learn about Czech grammar, whether Czech is their native or second language. The STYX system (Hladká and Kučera, 2008) is and electronic exercise book of Czech morphology and syntax based on the Prague Dependency Treebank (PDT).[2]

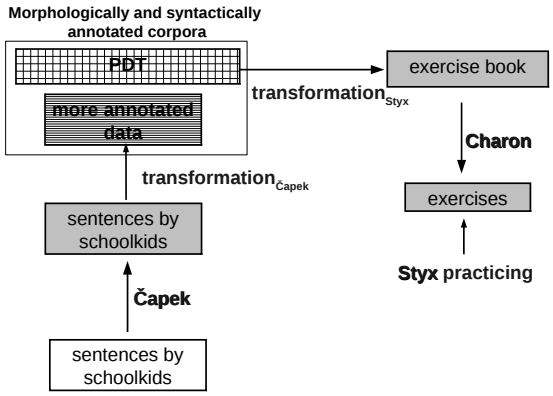

Figure 3: The STYX system

The STYX system consists of

- a database of 11,000 sentences with their morphological and syntactic analyses. The PDT data annotated on all annotation layers, i.e. on morphological, syntactic and semantic layers (49,442 sentences) constituted the candidate set of sentences from which the sentence database was composed. Then sentences from the candidate set unsuitable for schoolchildren because of too complex phenomena were discarded. This resulted in 11,718 sentences, the annotations of which were consequently transformed (see $transformation_{Styx}$) into the school annotation scheme - compare sentence[3] annotation in Figure 4 and Figure 5. In the school annotation scheme, a sentence is represented as a tree-like structure with the labeled nodes which corresponds to the representation of dependency-based syntax. Unlike a tree structure, this structure has no root node or, from another point of view, it has two roots: a subject and a predicate.
- a user interface
  - *Charon* to select sentences from the database, i. e. to compose the exercises. Because it would be time-consuming to go over all the 11,000 sentences, it is possible to select sentences with particular phenomena.
  - *Styx* to analyze the selected sentences both morphologically and syntactically, and to check the analyses. The morphological analysis comprises part-of-speech tag assignment; the syntactic analysis comprises tree-like structure building and syntactic tag assignment.

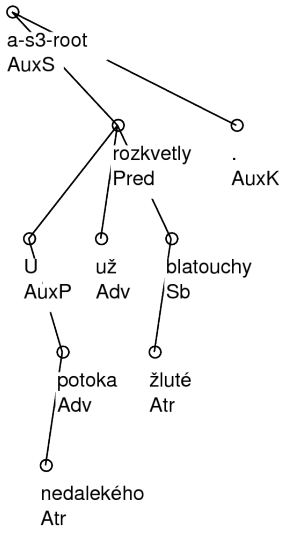

Figure 4: Sample of PDT annotation

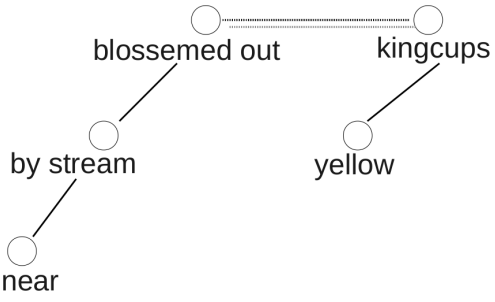

Figure 5: Sample of the Czech school annotation scheme

Another part of the Styx system is *Čapek*, an editor which can be used to annotate arbitrary sentences morphologically and syntactically following the school annotation scheme (see Figure 6). Once the sentences are read, they are tokenized and the user can proceed with practicing[4] morphology and syntax. The editor exists in two versions: (1) as a

[2] http://ufal.mff.cuni.cz/styx

[3] The white kingcups have blossomed out by near stream. *U [by] nedalekého [near] potoka [stream] už [already] rozkvetly [blossemed out] žluté [yellow] blatouchy [kingcups].*

[4] When speaking about language classes, we prefer to use the term *practicing* instead of *annotation*.

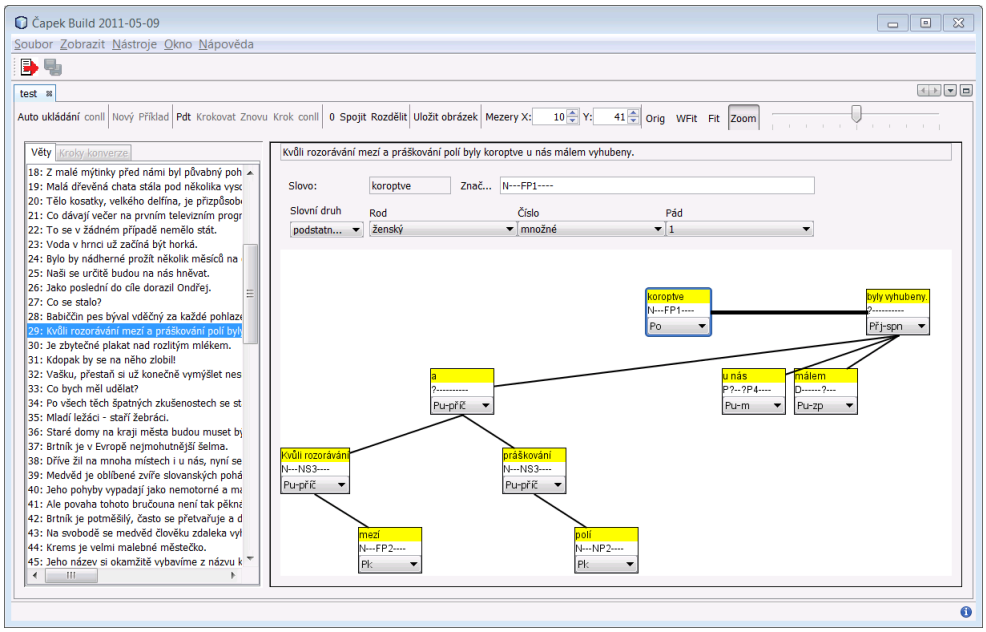

Figure 6: Screenshot of Čapek

desktop application, written in Java on top of the Netbeans Platform,[5] and (2) as an iOS application for iPad.

We offer the editor to teachers to use in their language classes instead of paper and pencil as they would normally do. They provide their own sentences and their pupils annotate them. We transform the result into our desired target annotation scheme and add it to an existing annotated corpus. This way we increase the volume of training data needed for supervised learning methods.

4. Czech textbook sentences and their processing

The *Čapek* editor is designed to enable annotation of arbitrary sentences. However, to evaluate the annotation with schoolchildren, we selected a sample of 100 sentences from Czech textbooks to serve as an annotation workbench. As can be seen in Figure 7, these sentences were annotated both manually and automatically:

- An automatic annotation by the perceptron-based parser (McDonald at al., 2005).
- A manual annotation by
  - an expert-linguist according to the PDT 2.0 annotation framework using the TrEd annotation editor.[6] We consider this annotation to be the gold-standard and we use it for the evaluation of transformation procedure $transformation_{Capek}$ (see Figure 3).
  - two teachers and two school children following the Czech school annotation framework using the Čapek editor.

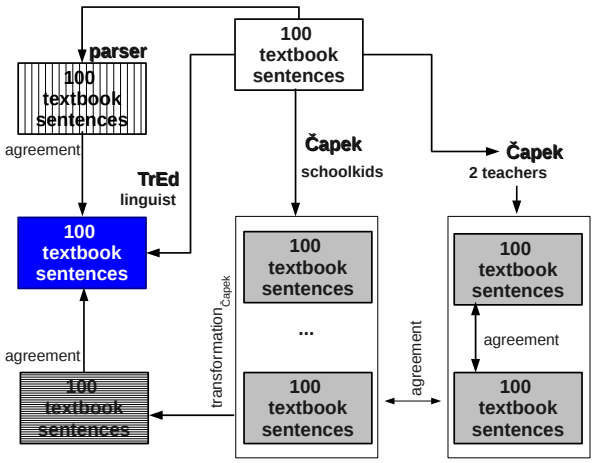

Figure 7: An annotation workbench

5. Transformation rules

As mentioned above, the school and PDT syntactic structures are slightly different. Therefore, we have developed a transformation procedure translating trees in the school system to the corresponding trees in the PDT one. The translation procedure consists of several steps, each focusing on a certain aspect. For example,

1. **Tokenization:** First, we have to do a different tokenization. The school structure works with words only and (mostly) ignores punctuation. The PDT system works with tokens, as it is usual in NLP.
2. **Subject depends on predicate:** We translate the structures treating subject and predicates as equal into structures where the (head of the) predicate is the root and the subject depends on it. This is done for both top-level and embedded clauses.

[5] http://platform.netbeans.org
[6] http://ufal.mff.cuni.cz/tred

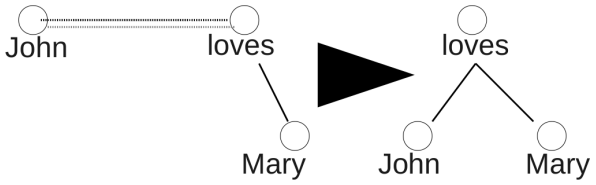

Figure 8: Subject-predicate transformation

3. **Tagset translation:** The syntactic tagset is translated using a simple dictionary. The PDT tagset contains tags not present in the school system, but they are not present at this stage, as they are used mainly with auxiliary words (see below) which do not have their nodes in the school system. These tags are introduced by later transformations as each of these phenomena is handled.

4. **Dropping PRO subject:** The school system introduces nodes for dropped subjects, which are common in Czech, while the PDT system does not represent them explicitly. We simply drop them from the structure.

5. **Splitting multi-token nodes**. As mentioned above, the school system operates on multi-word units. Auxiliary words do not have their own nodes, but instead they are considered to be some kind of inflection on content words. For example, the phrase *have been sleeping* and *sleeps* are both considered to be inflection of the verb *sleep* and thus both act as single units in the structure. Auxiliary verbs, all prepositions, modals, certain verbs such as *start*, *become*, etc. are all considered to be part of so called analytical morphology. We have to split these multi token units, as the PDT system treats all these words as separate items. This is done by a sequence of transformations:

   (a) **preposition + words:** The preposition is considered to be the head while the other word (usually a noun) is the dependent. All dependents of the unit will depend on the noun.

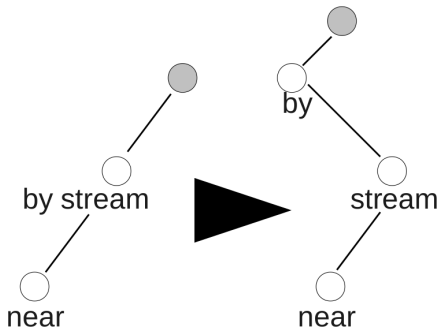

Figure 9: Splitting multi-token nodes with preposition

   (b) **reflexives:** As other Slavic languages, Czech uses reflexives extensively to express several functions (regular reflexive, passive, mutuality), in addition, reflexive-tantum verbs are always accompanied with a reflexive (*smát se* 'laugh' lit: 'laugh oneself'). These functions are distinguished in the PDT system. Currently we simply separate the reflexive pronoun from the verb and mark it as a part of a reflexive-tantum verb as this is the case in 65% of the cases (at least in the PDT data). In the future, we are planning to use VALLEX (Žabokrtský and Lopatková, 1972), a valency lexicon, to separate the reflexive-tantum and other uses. We can also use the label suggested by a parser or to train a dedicated classifier.

   (c) **'to be':** Forms of the verb *být* 'to be' are used in several complex verbal forms (future, past, conditional, nominal predicates). The PDT scheme distinguishes between auxiliary and non-auxiliary use, but since the distinction is rather subtle and the auxiliary use is by far the more common, we mark all such uses of the verb 'to be' as auxiliary.

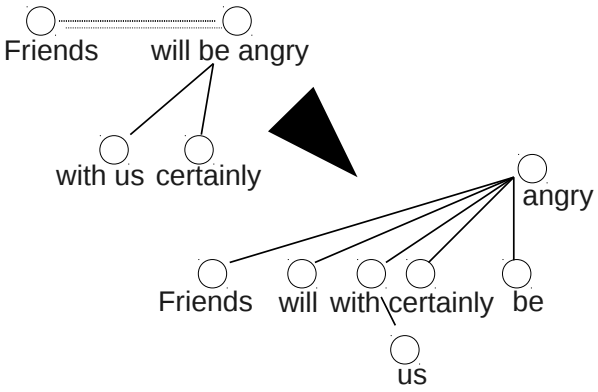

Figure 10: Splitting multi-token nodes with reflexive

   (d) Remaining multi-token units are arbitrarily split into a right branching sequence of nodes.

Examples of untreated phenomena:

1. All complex predicates beyond 'to be' + full verb (*stát se učitelem* 'become a teacher', *začít se třást* 'start to tremble', etc.) are handled by a very simple rule. First the predicate itself is split into right branching sequence of nodes, which is correct in the vast majority of cases. However, the main problem is in the distribution of dependents to the new nodes. In many cases, this cannot be done on purely syntactic grounds. For example, adverbs modifying the phrase *to start to shiver* can modify either the verb *start* or the verb *shiver*. Without considering semantics one cannot really decide which is the correct analysis as Figure 11 shows. Our procedure assigns all such dependents (except the subject) to the most embedded verb.

2. Coordination, apposition, insertions.

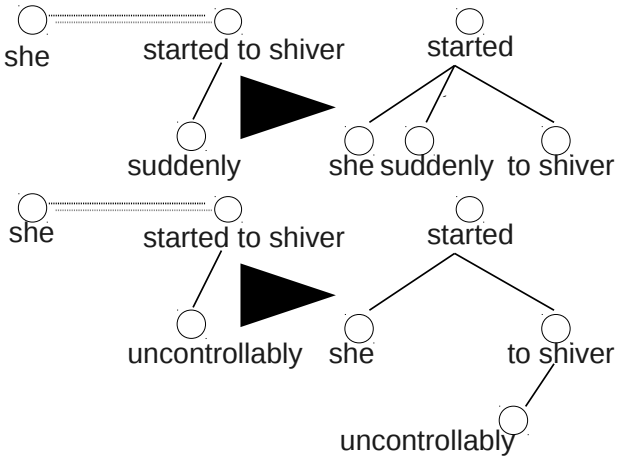

Figure 11: Adverb attachment uncertainty

All these translations can be performed within our editor – see Figure 12. The environment allows doing all transformations to be performed in a batch fashion for all sentences or step by step for individual sentence.

6. Conclusion

Using the annotation editor Čapek school children practice morphology and dependency-based syntax and consequently their annotation is transformed into the target annotation scheme. In our case, the object language is Czech and the target annotation scheme corresponds to the Prague Dependency Treebank annotation framework.

The transformation rules were designed to cover typical sentences practiced at Czech schools. We selected a representative sample of 100 sentences and this collection underwent both manual and automatic processing. We use the CoNLL 2007 *Unlabelled attachment score* (UAS) that measures the percentage of tokens with a correct/same head (i.e., only a correct/same dependency arc) as the agreement metric. The results are summarized as follows: UAS(parser *vs.* gold-standard) = 89%, UAS(teacher1 *vs.* teacher2) = 92%, UAS(children1 *vs.* children2) = 85%, and UAS(transformation of teacher1's annotation *vs.* gold-standard) = 81%.

Simply said, if we want to get more annotated data via the school annotation and its transformation, the parser's accuracy must be beaten. The accuracy of the 'annotation+transformation' procedure is significantly lower than the parser's accuracy. Since we have the annotation by only two teachers, two kids and a limited number of transformation rules, we cannot make final conclusions, we can make just positive and negative observations. While the parser accuracy is based on millions of statistics (mostly probabilities), the transformation's accuracy is a result of less than ten rules. We are aware of phenomena not covered by these rules and we believe that covering them will significantly improve the transformation accuracy.

However, we were quite surprised by the relatively low mutual agreement between teachers, despite the fact that they invested were annotating very carefully. Some of the discrepancies were due to errors, others were the result of different interpretation of the theory and/or individual sentences. For example, the teachers often differed whether they treated a complex predicate as a single unit or several units (is the verb *have* in *have an opportunity* an auxiliary or a full verb). It is a known fact that experts often disagree on annotation. However, in our case, they were not annotating the usual type of sentences found in newspaper-based corpora, but examples from a school textbook carefully selected (or in fact created) by the authors to correspond to the level of the syntactic theory taught in schools. The sentences are not simple, but they do not contain coordinations of unlikes, etc. either. Actually, if the teachers graded each other the way they do grade their pupils, the grades would definitely not be As.

On the other hand, we were surprised by the high mutual agreement between the school children, which was close to that of the teachers. However, we consider our subjects to be gifted students and realistically, we estimate mutual agreement for school children to be slightly below 80% in average.

We got a positive feedback from both students and their teachers, and several schools have expressed an interest to use our editor for class exercises and homeworks.

No matter how the national curricula organize language classes, we provide a crowdsourcing way how to enlarge amount of annotated data.

7. Acknowledgments

The authors wish to thank Lucie Medová for her active assistance in research on sentence diagramming at schools over the world. We also thank Marie Konárová for collecting the annotated data and to the teachers and school children for their time and effort spent with annotation.

This research was supported by the Grant Agency Czech Republic (projects ID: P406/10/P328 and P406/12/0658).

8. References

Luis von Ahn and Laura Dabbish. 2008. Designing Games with a Purpose. *Communications of the ACM*, vol. 51, No. 8, pp. 58–67.

Thorsten Brants, Ashok C. Popat, Peng Xu, Franz J. Och, Jeffrey Dean. 2007. Large Language Models in Machine Translation. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pp. 858–867. Prague, Czech Republic.

Barbora Hladká and Jiří Mírovský and Jan Kohout. 2011. An attractive game with the document: (im)possible? *The Prague Bulletin of Mathematical Linguistics*, pp. 5–26. No. 96.

Ryan McDonald and Fernando Pereira and Kiril Ribarov, and 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, Canada, pp. 523 – 530.

Barbora Hladká and Ondřej Kučera. 2008. An Annotated Corpus Outside Its Original Context: A Corpus-Based Exercise Book. In *Proceedings of the ACL-08: HLT*

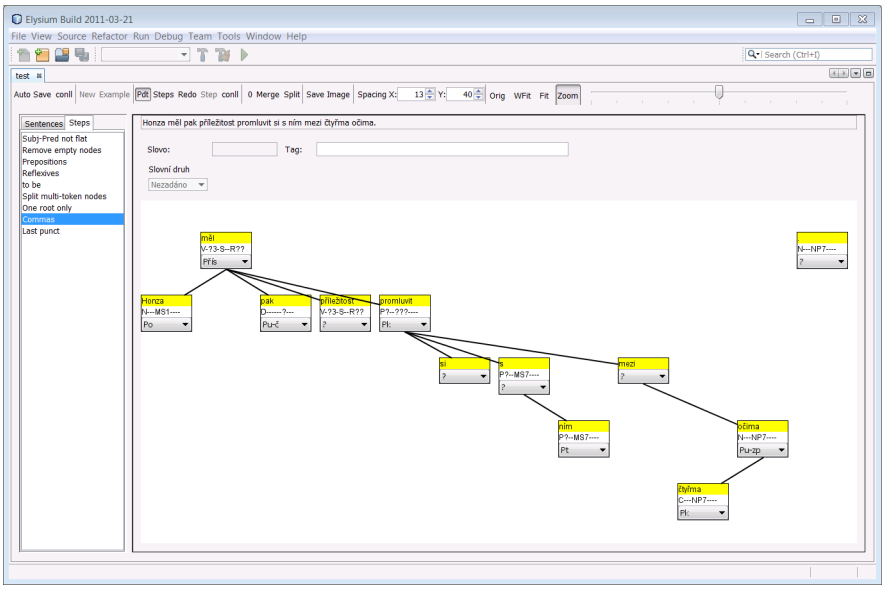

Figure 12: Transformation from the school system to the PDT system within the editor

Third Workshop on Innovative Use of NLP for Building Educational Applications, pp. 36-43, The Ohio State University, Columbus, Ohio, USA.

Brainerd Kellogg and Alonzo Reed. 1899. Title Higher Lessons in English. A work on English grammar and composition.

Richard E. Hudson. 1992. *Teaching Grammar: A Guide for the National Curriculum (Language in Education).*

Vladimír Šmilauer 1972. *Nauka o českém jazyku.* Praha:Státní pedagogické nakladatelství.

Zdeněk Žabokrtský and Markéta Lopatková. 2007. Valency Information in VALLEX 2.0: Logical Structure of the Lexicon. *The Prague Bulletin of Mathematical Linguistics*, pp. 41–60 No. 87