# Referee report on the PhD Thesis by Petr Horáček

Akihiro Yamamura, Akita University

## 1 Summary

The author studies rule-synchronization approach to formal languages and its application to natural language processing, in particular, to understand:

- Introduction of a rule-synchronized context-free grammar
- Analysis of generative power of synchronous grammars
- Introduction of a rule-restricted transducer
- Applications to natural language translation

He writes four papers concerning on his research theme.

1. "Regulated rewriting in natural language translation"
2. "Synchronous versions of regulated grammars: generative power and linguistic applications"
3. "Rule-restricted automaton-grammar transducers: power and linguistic applications"
4. "New grammar systems and their application perspectives"

Then he summarizes the obtained results in his thesis "SYNCHRONOUS FORMAL SYSTEMS BASED ON GRAMMARS AND TRANSDUCERS". I as a referee read the papers and the thesis and express my opinion.

In the first paper, the author introduces the concept of synchronous grammar based on linked rules and discusses synchronous scattered context grammar and synchronous matrix grammar. He also presents an application of the synchronous matrix grammar to Czech-Japanese translation.

Then he considers generating power of synchronous grammar in the second paper. He proves that $\mathscr{L}(RSCFG) = \mathscr{L}(MAT)$, that is, the family of languages generated by rule-synchronized context free grammars coincides with the family of languages generated by matrix grammar. This is surprising because the generating power of context free grammar with linked non-terminals does not exceed the one of context free grammar. Therefore, it is very interesting result and it makes sense to study synchronous grammar defined in the first and second paper where synchronization is given by linked rules. The author shows that $\mathscr{L}(SMAT) = \mathscr{L}(MAT)$ and $\mathscr{L}(SSCG) = \mathbf{RE}$ and presents an application to English-Japanese translation as well. Translation can be considered as a pair of sentences generated by applications of linked rules. Therefore the synchronization mechanism introduced in these papers can be applied to translation of natural languages. His approach seems to have advantage of handling inflections.

In the third paper, he develops his concept of synchronization to more complex cases such as rule-restricted transducers that resembles synchronous grammar in using linked rules although the transducer is more complicated than synchronous grammars. A rule-restricted transducer introduced in the third paper can be considered as synchronization of a finite automaton and context free grammar. It looks hard to deal with non-determinism in this scheme and so the author put leftmost restriction on derivation in a rule-restricted transducer. It looks too restrictive at first glance, however, some restrictions should be assumed. Many theoretical results on languages generated (or accepted), for example, $\mathscr{L}(RT)_2 = \mathscr{L}(MAT)$, are obtained. Transform of active voice — passive voice transform is also discussed in addition to transforming a sentence in some natural language into another language. He demonstrates translation among Czech, English and Japanese. From the standpoint of translation, the approach by rule-restricted transducer is more realistic than the one of synchronous grammar because a transducer yields one sentence to a given one while one needs to parse applications of rules in generation process to produce a translation in a synchronous grammar approach. The fourth paper is a long summary of the papers above.

The theoretical results investigating generating power of introduced schemes like synchronous grammars and rule-restricted transducers are of great significance. The author shows novelty and originality. I believe that the results in the four papers and the thesis contribute to the research community on both formal language theory and natural language processing significantly.

## 2 Organization

The thesis consists of the following three chapters.

Chapter 1 explains motivation of the research and gives necessary terminology from mathematics. Computational linguistics is also discussed to some extent.

Chapter 2 presents synchronous mechanisms using linked rules: rule-restricted context free grammars, synchronous scattered context grammars, synchronous matrix grammars and rule-restricted transducers. The languages generated (or accepted) are discussed in depth.

Chapter 3 presents applications to natural language processing such as translations. Then the author Discusses future research and concludes the work.

## 3 Main Achieved Results

The author invents synchronization of language (generating or acceptance) on schemes using linked rules. The idea is novel and deeper than the previous approach using linked non-terminals, where generating power of does not increase over context free grammar. From theoretical point of view, comparison of generating power of the introduced schemes and the traditional grammars and automata are of great significance. For example, he show that the family of languages generated by rule-synchronized context free grammars coincides with the family of languages generated by matrix grammar. Similar technique is applied to the other investigations. Furthermore, his effort to apply to natural language processing should not be neglected. Natural language processing area is vivid and developing fast, however, the approach by the author is quite unique and completely new and so it impacts on NLP community.

# 4 Drawbacks

The work has no significant drawbacks. As minor drawbacks, I consider the following two aspects of the papers.

(1) Linked rules in the two languages of a synchronous grammar corresponds one-to-one in the paper "Regulated rewriting in natural language translation", that is, there is a bijection of the set of rules in $G_I$ onto the set of rules in $G_O$ . On the other hand, it is not assumed there exists a bijection between the set of rules in the synchronous grammar in the following papers and the thesis. Strictly speaking, the schemes introduced in " Regulated rewriting in natural language translation " are different from the ones in the other papers. I think it is necessary to allow one rule in one side corresponds to many rules in the other side to prove Theorem 1 in the paper "Synchronous versions of regulated grammars: generative power and linguistic applications". It is better to explain why the definition is changed afterward. I am even interested in the case that rules in both sides corresponds one-to-one. Does it have the same generating power in such a case?

(2) The author discusses applications of the proposed schemes to natural language processing such as translation by demonstrating synchronous grammar or a rule-restricted transducer. It is desired to explain shortcomings of existing research on NLP and an advantage of the proposed scheme. Some background is given in Section 3.2 in the thesis. I am interested in merits and demerits of existing approaches.

# 5 Open Problems

Although the thesis has achieved many important results, there still remain open questions related to this work. I am interested in the following. Have you studied any of them?

(1) I wonder if synchronization of more than two grammars can increase language generation power as the author mentions as one of further research topics in his thesis. This is related to multi-linguistic translation such as English-Czech-Japanese simultaneous translation and may give more insight into natural language structure. It may possible to invent a transducer of one input sentence and two output sentences in the different natural languages.

(2) It may be possible to conclude some results on the structure of natural languages using the results obtained. For example, the existence of translation between two languages implies either of natural language cannot be modeled by a context free grammar because of the result $\mathscr{L}(RSCFG) = \mathscr{L}(MAT)$. Conversely, if two natural languages are modeled by context free languages, they cannot be translated by the scheme introduced. This suggests that natural languages have more complicated structure than context free languages, or translation exceeds formal language approach. You may be able to conclude some results on natural languages using the theoretical results.

(3) As I mentioned above, it may be interesting to study restricted case that rules in both sides corresponds one-to-one. Probably, it has the same language generating power although I do not know whether or not it is correct.

## 6 Suggested Corrections of Minor Mistakes

I suggest a minor correction.

Page 35. line 16 of the thesis:
The definition of $\mathscr{L}(RSCFG)$ is not clear enough. What does it mean by the phrase "as their input language" ? I understood that $\mathscr{L}(RSCFG)$ consists of input languages of a $RSCFG$.

## 7 Conclusion

T strongly believes the author's approach presents a new perspective of synchronization of language generation and its applications to natural language processing open new research areas. The thesis is well written and the author achieves the standard of research community and I believe that the author will be an active researcher in formal language theory. Therefore,

**I strongly recommend the Ph.D. thesis by Petr Horáček for the defense.**

Professor Akihiro Yamamura, Ph.D.
Akita University
Japan