EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Philosophische Fakultät

Slavisches Seminar

Prof. Dr. Tilman Berger
Slavische Sprachwissenschaft

Slavisches Seminar
Wilhelmstraße 50
72074 Tübingen
Telefon +49 7071 29-76733
Telefax +49 7071 29-5924
tberger@uni-tuebingen.de
www.slavistik.uni-tuebingen.de

Universität Tübingen · Wilhelmstr. 50 · 72074 Tübingen

Oddělení pro vědu a zahraniční styky
Matematicko-fyzikální fakulta Univerzity Karlovy
Ke Karlovu 2027/3
12116 Praha 2

Česká republika

Tübingen, May 31, 2024

Review of the Habilitation Thesis "Cross-Language Harmonization of Linguistic Resources" of Dr. Daniel Zeman

Before I give my review, I would like to start with a short personal comment and recall my first meeting with Dr. Zeman, which took place more than 20 years ago at ÚFAL. I do not remember the exact year, but I remember that Dr. Zeman talked about the problem that different national corpora use different tag sets and that this is the reason why it is difficult to compare corpus data from different languages in a reasonable way. And he outlined his idea of defining a universal tagset that would allow standardized searching.

After that we met only once and had no further contact. But I learned that he had realized his idea, not only for the field of morphology, but also for syntax, and I was fascinated by these works. Today I am very happy to have the opportunity to write a review of some of his works, namely for his habilitation.

I will do this from the perspective of a linguist who has experience with corpora and also with typologically different languages, and I hope to be able to contribute to the review of the work from this perspective. On the other hand, I can only comment on the computational linguistic aspects to a very limited extent.

Dr. Zeman's work is divided into three major parts, namely work on "low-resource languages", the "harmonization of morphological annotations" and the "harmonization of syntactic annotations". In the following, I will comment on each part and discuss the articles that belong to it.

The first part of the thesis on "low-resource languages" deals with something that a traditional linguist cannot imagine, or that many would consider impossible. It asserts (and to some extent proves) that it is possible to apply techniques such as parsing to texts in low-resource languages without having a lot of data about these languages. This is done either by delexicalized parsing (the parser then analyses only the morphological tags) or by ana-

lyzing parallel corpora. Dr. Zeman himself has mainly worked with the first method, which he discusses in more detail in the 2008 paper “Cross-Language Parser Adaptation between Related Languages” he co-authored with Philip Resnik, but he is also involved in the evaluation and comparison of both methods. And the results are very impressive indeed, giving hope for further progress in the coming years. So far, the method has been applied to several European and some Indian languages. It would be interesting to see results from typologically different languages from all over the world.

The second part of the thesis deals with the harmonization of morphological annotations, which is a central prerequisite for the procedures described in the first part (and at the same time the topic that I found so interesting many years ago). As a Slavist who has been working with corpora for many years, I am familiar with the difficult (and regrettable) starting position, because each national linguistics has first developed its own tagsets, often based on native traditions, and a comparative search was not really possible. Dr Zeman has done a great deal of work in this area over the last sixteen years, particularly with the development of tagset drivers for converting from one tagset to another, and then with the development of Interset, a sort of universal tagset for 40 languages. This topic is very succinctly described in the 2008 paper “Reusable Tagset Conversion Using Tagset Drivers”. In a further step, the universal tagset was integrated into the “Universal Dependencies” project, which also includes syntax and which is described in the 2021 paper of the same name.

The task of unifying tagsets is not as simple as a naïve observer might initially assume. Grammatical categories such as genus can be assigned to different word types in different languages and can also differ in terms of their assignment to inflection or derivation. When merging tagsets, decisions have to be made here that should also be justified. – An even more difficult problem are features that are tagged more than once in a word (the author deals with this under the heading “Layered Features”). In solving this problem, however, Dr. Zeman shows his excellent linguistic knowledge. He argues – e.g. on the categories of 3rd person pronouns in Croatian – with a good knowledge of the linguistic facts and can then also propose a very plausible solution.

I would like to mention another very important aspect here, namely the fact that all this work has been done in close collaboration with other researchers around the world and in a constant dialogue within the discipline. This is where the term “universal properties” takes on another very important meaning. Nowadays, it is no longer enough to work alone or in a small group to find a solution; it has to be done in collaboration with people working on similar topics. Dr Zeman has an excellent network in the international community and can meet this requirement in the best possible way.

The third part of the thesis deals with the harmonization of syntactic annotations. This is an even more challenging task than the first two, because the existing systems differ considerably more than in morphology. This major undertaking is presented in great detail and with many diagrams in the papers “HamleDT: Harmonized Multi-language Dependency Treebank” (2014) and “Universal Dependencies” (2021) and applied to a large number of languages. I won't go into any more detail on this and will limit myself to saying that I can understand all the results and that I am fascinated by what is possible in this field.

The last two papers extend the arc even further. The first paper “CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies” deals with the whole way

from an unanalyzed text to a description of its structure, and is also the result of a close international collaboration with partners from several countries. It presents further refinements and evaluates them in detail. The second paper “Towards Deep Universal Dependencies” (2019) sketches the first steps of a new journey towards deep syntax and semantics. Both papers show how the author is constantly developing his ideas in new directions, but I will not discuss them in detail here.

The analysis of the thesis by the Turnitin system shows some matches to project and publisher websites where Dr Zeman's work has been published. Such matches are normal and occur wherever an author writes frequently on similar topics.

Overall, the author has produced an impressive piece of work. He presents unified and universal solutions in areas where several methods have previously competed, and he has also succeeded in applying these methods to low-resource languages, something that many people (myself included) did not think possible. I have no hesitation in confirming that this thesis meets and exceeds the standard requirements for a habilitation. I strongly recommend that the habilitation process be continued.

Tübingen, May 31 2024

Prof. Dr. Tilman Berger