Posudek diplomové práce

Matematicko-fyzikální fakulta Univerzity Karlovy

Autor práce Christian Cayralat
Název práce Orthography Standardization in Arabic Dialects
Rok odevzdání 2021
Studijní program Informatika **Studijní obor** Matematická lingvistika

Autor posudku Pavel Straňák **Role** oponent
Pracoviště Ústav formální a aplikované lingvistiky, MFF UK

Text posudku:

The topic of the presented thesis: standardisation of the orthography in Arabic dialects, i.e. Arabic language normally used in all Arabic-speaking countries, is a very interesting one. It is also very broad, which results in the fact that the thesis touches upon many sub-topics that it doesn't solve. I don't see this necessarily as a problem, as explained below.

The author describes the situation with Arabic language, lack of any standardisation for the regional and national varieties (called dialects), available datasets, and the results of the situation for processing Arabic, especially informal, using NLP methods. The author uses his Lebanese background and concentrates on the Lebanese Arabic standardization.

The thesis is written in very good English with only occasional small errors. It is also very thoroughly typeset, which is not trivial, especially given the number of Arabic examples, and has exemplary use of cross-references. All in all, it is on a high level formally, which strongly improves readability of the complex text.

Complexity of the thesis comes from the fact that it tackles a complex area and, seeks not to beat a concrete result, but to see the whole task of orthography standardization better. I believe that the goal was if not achieved, then good foundation for it was lied down with the thesis.

The thesis is explorative, and the exploration is scientifically thorough. From the review of the current literature, setup of the first experiments, analysis of their problems, including the parts that the author was not able to solve in his thesis, or he didn't have time to explore further. When the experiments led him to the need to create a dataset, he set up a sensible annotation workflow, including a non-trivial annotation tool, and created what I believe is a valuable and in some aspects – correctly identified in the thesis – unique corpus. The fact that the extent of the corpus is limited and some annotation needs further review doesn't take away the value of the achievement.

The new dataset is finally used to complement the first series of experiments run on an existing corpus by a set run on this new data with different characteristics, which can be also studied from the comparison of these two sets of experiments. The attempt to systematically classify the morphonological differences of the Lebanese dialect from the Modern Standard Arabic during annotation, and later trying to train a classifier to identify these types of differences during standardisation presents also a substantial linguistic work, de facto partially but systematically describing Lebanese Arabic grammar specifics. I might have some reservations to the terminology used and invented, but those are absolutely minor, compared to the extent of the work.

In my opinion the author demonstrated his ability to work systematically and creatively in the field of mathematical linguistics in both the computational and linguistic aspects in an unusually equal measure. The thesis provides a valuable basis for future research.

Práci doporučuji k obhajobě.
Práci nenavrhuji na zvláštní ocenění.

V Praze dne 31. srpna 2021

Podpis: