Towards the automatic detection of syntactic differences

Kroon, M.S.

Towards the Automatic Detection of Syntactic Differences

Natural language syntax is the system of combinatorial rules that builds complex hierarchical structures, i.e. phrases and clauses, out of individual words and morphemes. The insight that the words of a sentence are organized both linearly and hierarchically, i.e. as phrases that contain phrases that contain phrases, is central in modern linguistics. The field of theoretical comparative syntactic research aims to identify the range and limits of syntactic variation between natural languages by comparing their structures and describing the syntactic similarities and differences, and to capture them in a cross-linguistic formal theory.

This dissertation centers around the question whether syntactic differences between languages can be detected automatically, and if so, how. With the enormous number of natural languages and dialects, the very high level of variation they exhibit between one another, and the technically infinite number of possible sentences per language or dialect, systematic manual comparison is a hugely daunting task. The field would therefore significantly benefit from the (partial) automatization of the process, as it would increase the scale, speed, systematicity and reproducibility of research.

Over the course of five chapters it is shown through case studies involving English, Dutch, German, Czech and Hungarian that correct hypotheses on syntactic differences between languages can be generated automatically from parallel corpora through the use of the minimum description length principle, counting mismatches between part-of-speech pattern occurrences, word alignment and mapping annotation from an annotated language onto another unannotated language. The tools developed for the purposes of this research work well and can aid a linguist significantly in their search for differences or similarities, but do not replace the human researcher.

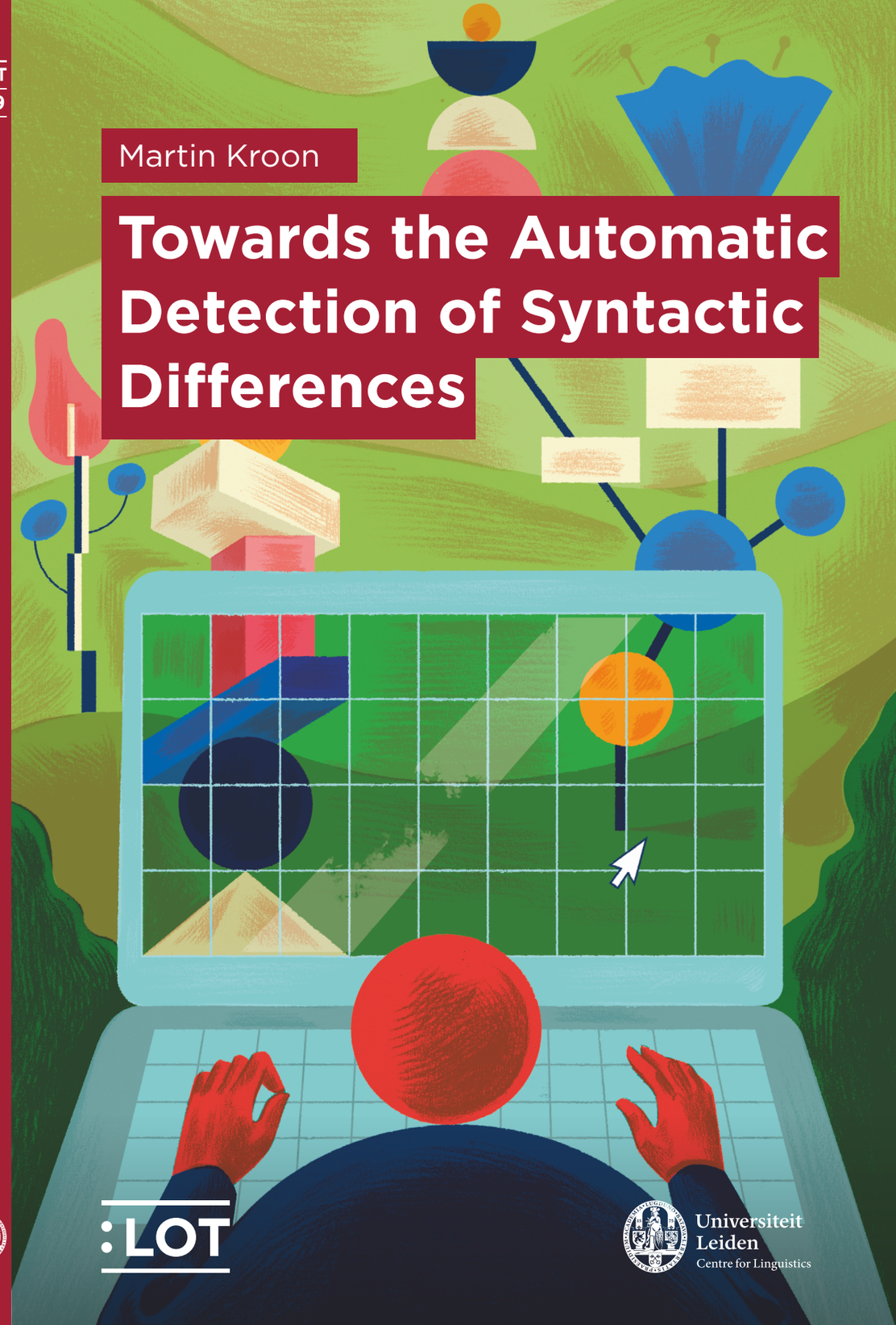
ISBN 978-94-6093-414-8
DOI https://dx.medra.org/10.48273/LOT0629

:LOT
Netherlands Graduate School of Linguistics
Landelijke Onderzoekschool Taalwetenschap

:LOT
629

Martin Kroon

Towards the Automatic Detection of Syntactic Differences



Martin Kroon

Towards the Automatic Detection of Syntactic Differences

:LOT

Universiteit Leiden
Centre for Linguistics