

In the field of Natural Language Processing, word-formation is under-resourced comparing to inflectional morphology. Moreover, the existing resources capturing word-formation differ in many aspects. This thesis aims to review existing language resources for word-formation across languages and to unify them to a common data structure and file format. Basic notions of word-formation are followed by a review of existing language resources and their comparison in both quantitative and qualitative aspects. In the core part of the thesis, the harmonisation process is presented. Design decisions on the unification procedure are presented, and the selection of the resources to unify is described. The resources are unified to the rooted tree data structure and stored in a lexeme-based file format, which is already used in DeriNet 2.0. The procedure applies supervised machine learning model and the Maximum Spanning Tree algorithm. While the model scores word-formation relations, the MST algorithm uses the scores for identifying the rooted tree structure in each word-formation family. The resulting collection of harmonised resources covering 20 European languages was published under the title 'Universal Derivations' (UDer).