Tagungsbericht / Conference Report

"DeriMo 2019. Second Workshop on Resources and Tools for Derivational Morphology" 19–20 September 2019; Prague, Czech Republic (Petra Steiner)

DeriMo is an international meeting dealing with derivational morphology from the perspective of data analysis. Its second edition DeriMo 2019 was held at the Faculty of Mathematics and Physics, Charles University in Prague. The local organizers are researchers of the Institute of Formal and Applied Linguistics (ÚFAL = Ústav Formální a Aplikované Linguistiky) at the Computer Science School of the Faculty of Mathematics and Physics. Chairs of the program committee were Magda Ševčíková (ÚFAL), Zdeněk Žabokrtský (ÚFAL), Eleonora Litta Modignani Picozzi (CIRCSE, Università Cattolica del Sacro Cuore, Milan), and Marco Passarotti (CIRCSE, Università Cattolica del Sacro Cuore, Milan).

ÚFAL with its strong and long tradition from Prague School and functionalism is dedicated to approaches of Natural Language Processing based on theoretical frameworks such as Dependency Grammar or Quantitative Linguistics. This institutional background and the work on morphological tools and dictionaries were the pivots of the workshop. One of the latest developments at ÚFAL is DeriNet, a network of word-formation for the Czech language. It is one of 22 databases with derivational information listed by Kyánek (2018). 17 of these and many other derivational resources were dealt with in the presentations. The contributions presented recent developments in word-formation resources and tools, research by using morphological data, and theoretical accounts and descriptions of word-formation processes.

This workshop gathered researchers from typology, computational linguistics, classical philology, and grammatical description frameworks. The languages covered included Bulgarian, Croatian, Czech, English, Estonian, Farsi, French, German, Japanese, Latin, Polish, Portuguese, and Spanish, besides many others in cross-linguistic studies. These two days' talks dealt with a wealth of morphological data, tools, and methods not just for applications in natural language processing but also for research in morphology, typology, semantics, quantitative linguistics, language teaching, or speech therapy. Researchers with a need for high-quality morphological data should have a look at this workshop's proceedings to get an overview of

© 2020 Petra Steiner https://doi.org/10.3726/zwjw.2020.01.03

the state of the art. Most of the presented sources are publicly available. The workshops comprised 14 contributions of 24 authors. The papers, slides and one handout are available at the conference website: http://ufal.mff.cuni.cz/derimo2019/.

The first of two invited talks, *Cross-linguistic research into derivational networks* by Lívia Körtvélyessy (P.J.Šafárik University, Košice, Slovakia), introduced to derivational networks that are multi-dimensional description frameworks for derivational processes. Körtvélyessy's project aims at comparing the derivational complexity of 40 European languages. Chains of consecutive derivations yield different levels of forms and semantic categories. The study aims at characterizing their derivational and paradigmatic properties by defining measures: The paradigmatic capacity of a word is characterized by the number of its potential derivations, the number of consecutive morphs is defined as the order. Consecutive semantic categories show the semantic shifts of the derivational processes. Another characteristic measure is the Maximum Derivational Network (MDN), which is the sum total of maximums of the derivatives of all semantic categories. The MDN is a measure for the derivational potential. The saturation value is the ratio of the number of actual derivatives and the MDN; it describes the realization of the potential. These values were calculated for each order and the sums of all orders. Only six of the 40 languages reach the maximum order of 5 for all parts of speech, among them Czech with its complex derivational morphology. German comes only to an order of 3, with a high saturation value of 45.45 compared to the Czech value of 27.91. Furthermore, the investigation shows that the choice of semantic categories depends on the order and some language-specific blocking effects. This ongoing research gives new perspectives on cross-linguistic patterns of word-formation.

Daniele Sanacore, Nabil Hathout, and Fiammetta Namer (University of Toulouse/University of Lorraine, France) presented *Semantic descriptions of French derivational relations in a families-and-paradigms framework*. They link data of Démonette, a derivational database for French (Hathout & Namer 2014), with the semantic database FrameNet by building frame-like structures for derivational paradigms. By this, they show that at least some FrameNet frames can be adapted for describing morphosemantic relations of derivational families.

Lucie Pultrová's (Charles University, Prague) contribution *Correlation between the gradability of Latin adjectives and the ability to form qualitative abstract nouns* provides strong evidence for the correlation between the gradability and the ability of adjectives to form abstract nouns by investigating the suffixes *-tas*, *-itia*, *-tudo* and *-ia* in the database Bibliotheca Teubneriana Latina III.

In *The Treatment of Word Formation in the LiLa Knowledge Base of Linguistic Resources for Latin*, Eleonora Litta, Marco Passarotti, and Francesco Mambrini (Università Cattolica del Sacro, Milan, Italy) described the union of linguistic resources and tools for Latin within a new knowledge base. Here, all entries in lexical resources and corpus tokens that refer to the same lemma are linked within the Word Formation Latin Lexicon.

In her talk *Combining Data-Intense and Compute-Intense Methods for Fine-Grained Morphological Analyses*, Petra Steiner (Friedrich Schiller University of Jena, Germany) reported on a hybrid approach for analyzing German derivational and compositional morphology. This combines deep-level morphological structures of a database derived from GermaNet and CELEX, the results of a conventional word segmenter, and frequencies of constituents within contexts drawn from a large Wikipedia corpus.

The talk by Klára Osolsobě (Masaryk University, Brno) on *The Tagged Corpus (SYN2010) as a Help and a Pitfall in the Word-formation Research* describes the problems the SYN2010 corpus caused for the Dictionary of Affixes used in Czech (Slovník afixů užívaných v češtině; SAUČ). Especially graphical variants and overgeneration of morphological analyses lead to wrong entries.

In *Attempting to separate inflection and derivation using vector space representations*, Rudolf Rosa, and Zdeněk Žabokrtský (Charles University) explored the possibility of automatically classifying inflected and derived word forms. Their work builds on the working assumption of a clear-cut boundary between these categories and the contextual notion of meaning. The experiments used word embedding similarities, edit distance measures, and the combinations of both. The main hypotheses were that inflected forms are more similar and semantically closer than derived forms. The test data consists of 69,743 word forms of derivational families with 4,514 lemmas taken from DeriNet. With F1 scores around 40%, close inflections and derivations could be distinguished from a large set of different forms. A further analysis of particularly difficult cases shows that cos distance clearly separates inflected from derived forms.

The talk *Redesign of the Croatian derivational lexicon (CroDeriV)* by Matea Filko, Krešimir Šojat, and Vanja Štefanec (University of Zagreb, Croatia) describes their diligent lexicographic work in developing and enriching CroDeriV, which they expanded by 6,000 nouns and 1,000 adjectives. The morphological structures were manually analyzed by a two-level morphology, and the allomorphs of stems and affixes annotated. Other information was automatically determined, such as the base or root of the word, and the type of the word-formation process.

The last talk of the first day was from Jonáš Vidra, Zdeněk Žabokrtský, Magda Ševčíková, and Lukáš Kyjánek (Charles University). In *DeriNet 2.0:*

Towards an All-in-One Word-Formation Resource, the new annotation scheme of the database was introduced: a new file format permits storing new features and detailed description: especially compounds can now be described as entities with two parents. Furthermore, information about morphological categories such as aspect, gender and animacy, five semantic labels (diminutive, possessive, female, iterative, and aspect), and pseudo morphemes and bound root morphemes can be added to the knowledge base.

The second day began with the second invited talk by Fiammetta Namer and Nabil Hathout (Nancy/Toulouse), *ParaDis and Démonette, From Theory to Resources for Derivational Paradigms*. Following the prototype version Démonettev1, Démonettev2 implements ParaDis ("Paradigms vs Discrepancies"), a paradigmatic model for representing morphologically complex lexical units by merging principles from lexeme-based and paradigm-based approaches. This permits aligning lexemes to sets of derivational families by multiple connections even if the relations of their formal and semantic levels are not isomorphic, as for the affix replacement for *banque* ('bankN'), *bancaire* ('bankAdj'), and *interbancaire* ('interbankAdj'). The entries are sets of relations between lexemes with annotations.

In *Building a Morphological Network for Persian on Top of a Morpheme-Segmented Lexicon*, Hamid Haghdoost, Ebrahim Ansari, Zdeněk Žabokrtský and Mahshid Nikravesh (Charles University, Prague) presented two novelties: The first is a new large segmented morphological lexicon of Persian with 45,000 entries. As Persian is a morphologically rich language with many derivational affixes, this required concerted work of manual annotation and correction of the preprocessed word lists. Secondly, the word segmentation lexicon was used for building a morphological network analogous to DeriNet – DeriNet.FA. The authors experimented with weaving new and unprocessed words into the network by combining the morphological tool Morfessor with a recursive algorithm.

Lukáš Kyjánek, Zdeněk Žabokrtský, Magda Ševčíková and Jonáš Vidra (Charles University Prague) described the development of a multilingual derivational database in their talk *Universal Derivations Kickoff: A Collection of Harmonized Derivational Resources for Eleven Languages*. The collection under the title Universal Derivations (UDer 0.5) comprises a harmonization of the data from Démonette, DeriNet, DeriNet.ES, DeriNet.FA, DErivBase, English WordNet, EstWordNet, FinnWordNet, NomLexPT, The Polish Word-Formation Network, and Word Formation Latin. The different schemas are unified as rooted trees under DeriNet 2.0 format, other structures are represented in secondary sets. UDer combines networks of different quality and depth. For instance, DErivBase 2.0 has an average number of nodes of 1.2 per tree compared to 4.7 for DeriNet. Most derivational trees of this German part consist of singleton nodes.

Christian Curtis's (University of Washington, USA) work on *A Parametric Approach to Implemented Analyses: Valence-changing Morphology in the LinGO Grammar Matrix* is rooted in the Head-Driven Phrase Structure Grammar (HPSG). The object of investigation were verbal derivations that alter the argument structure of verbs by increasing or decreasing their valency. Curtis augmented the HPSG rule system by attaching valency-changing operations to lexical rule types. Examples and evaluation were among others on Mayan and Bantu languages, Javanese, Lakota, Japanese, and Hungarian.

The final talk of the workshop *Grammaticalization in Derivational Morphology: Verification of the Process by Innovative Derivatives* was presented by Junya Morita (Kinjo Gakuin University, Nagoya, Japan). It dealt with a contrastive investigation of meaning shifts of Japanese and English affixes such as *-er* and *-ee*. Objects were hapax legomena and other nonce words from BNC and BCCWJ (Balanced Corpus of Contemporary Written Japanese). While English derivational suffixes are ambiguous concerning the semantic roles of agent and instrument, their Japanese counterpart are not.

Kyjánek, Lukáš. 2018. Morphological Resources of Derivational Word-Formation Relations. Techn. Report. Prague: Charles University. ÚFAL. TR-2018-61.

Dr. Petra C. Steiner
Friedrich-Schiller-Universität Jena
Faculty of Arts
Institute of German Linguistics
Fürstengraben 30
07743 Jena / Germany
petra.steiner@uni-jena.de