▶ OntoLex-Morph: Morphology for the Web of Data

Christian Chiarcos,
Maxim Ionov,
Applied Computational Linguistics, Goethe University Frankfurt, Frankfurt am Main, Germany
Katerina Gkirtzou,
Institute for Language and Speech Processing, Athena Research Center, Athens, Greece
Anas Fahad Khan,
Istituto di Linguistica Computazionale "A. Zampolli", Consiglio Nazionale delle Ricerche, Italy
Penny Labropoulou,
Institute for Language and Speech Processing, Athena Research Center, Athens, Greece
Marco Passarotti,
Matteo Pellegrini,
CIRCSE Research Centre, Università Cattolica del Sacro Cuore, Milan, Italy

Purpose: OntoLex-Lemon is a widely used community standard for publishing lexical resources in machine-readable form, and is in fact the predominant RDF vocabulary for this purpose. With the growing popularity and increasing adoption of this model for applications in both language technology and lexicography, a number of new modules have been developed in the past year to complement the OntoLex core vocabulary and its lexicographic follow up, lexicog. In this paper, we describe the current status of the development of the OntoLex-Morph vocabulary.

Design/methodology/approach: (1) Describe and motivate OntoLex-Lemon, (2) document its shortcomings with respect to morphology, (3) describe OntoLex-Morph, and (4) illustrate its application for modelling (a) morphological information from traditional print dictionaries, (b) inflection tables from grammar books, (c) inflection and word formation information in machine-readable dictionaries and computational lexicons, (d) morpheme inventories, (e) dictionaries used for morphological generation, (f) morphological generation rules for different languages.

Findings: We demonstrate that OntoLex-Morph is applicable to all these domains for modern and historical languages *on real-world data*.

Research limitations/implications: We show applicability across a wide range of requirements. However, this has so far been applied to well-researched and well-resourced languages, mostly from Europe. OntoLex-Morph has been designed to accommodate typologically different languages (e.g., agglutinative languages), but a large-scale

application to typologically diverse languages has not yet been demonstrated on a broad scale, due to the lack of resources for many morphologically rich languages. So far, examples considered from these languages are artificial and taken from books rather than real-world language resources, as these languages are notoriously under-resourced.

Practical implications: We consider the vocabulary mature enough to codify it into a W3C vocabulary. This talk aims to elicit a final round of feedback from the wider community before we begin to form it into a W3C Community Report. Once published, it will become an integral part of OntoLex and become a stable standard.

Originality/Value: Moderate originality, as we do not describe novel findings, but the consolidation of a model that has been under development for over 5 years. High value, as it marks the conclusion of the modelling process. Before the publication of the community report itself, a publication originating from this talk is likely to serve as a reference publication for modelling morphological resources with RDF and/or LLOD technology. It supersedes all prior publications on the topic.

Keywords: morphology, machine-readable dictionaries, OntoLex, standard development.

Research type: Conceptual paper