CCURL 2014
LREC 2014
Reykjavik
Workshop on "Collaboration and Computing
for Under-Resourced Languages in the Linked Open Data Era"

Reykjavík (Iceland), 26th May 2014

Session 1

11:00-13:00

Chairperson: Joseph Mariani

11:00-11:30

The Multilingual GRUG Parallel Treebank – Syntactic Annotation for Under-Resourced Languages

Oleg Kapanadze

In this paper, we describe outcomes of an undertaking on building Treebanks for underresourced languages Georgian, Russian, Ukrainian, and German - one of the "major" languages in the NLT world (Hence, the treebank 's name – GRUG). The monolingual parallel sentences in four languages were syntactically annotated manually using the Synpathy tool. The tagsets follow an adapted version of the German TIGER guidelines with necessary changes relevant for the Georgian, the Russian and the Ukrainian languages grammar formal description. An output of the monolingual syntactic annotation is in the TIGER-XML format. Alignment of monolingual repository into the bilingual Treebanks was done by the Stockholm TreeAligner software. The parallel treebank resources developed in the GRUG project can be viewed at the URL of Saarland and Bergen Universities: http://fedora.clarin-d.uni-saarland.de/grug/ , http://clarino.uib.no/iness.

11:30-12:00

Multilingual Lexicography with a Focus on Less-Resourced Languages: Data Mining, Expert Input, Crowdsourcing, and Gamification

Martin Benjamin, Paula Radetzky

This paper looks at the challenges that the Kamusi Project faces for acquiring open lexical data for less-resourced languages (LRLs), of a range, depth, and quality that can be useful within Human

Language Technology (HLT). These challenges include accessing and reforming existing lexicons into interoperable data, recruiting language specialists and citizen linguists, and obtaining large volumes of quality input from the crowd. We introduce our crowdsourcing model, specifically (1) motivating participation using a “play to pay” system, games, social rewards, and material prizes; (2) steering the crowd to contribute structured and reliable data via targeted questions; and (3) evaluating participants’ input through crowd validation and statistical analysis to ensure that only trust-worthy material is incorporated into Kamusi’s master database. We discuss the mobile application Kamusi has developed for crowd participation that elicits high-quality structured data directly from each language’s speakers through narrow questions that can be answered with a minimum of time and effort. Through the integration of existing lexicons, expert input, and innovative methods of acquiring knowledge from the crowd, an accurate and reliable multilingual dictionary with a focus on LRLs will grow and become available as a free public resource.

12:00-12:30

Towards a Unified Approach for Publishing Regional and Historical Language Resources on the Linked Data Framework

Thierry Declerck, Eveline Wandl-Vogt, Karlheinz Mörth, Claudia Resch

We describe actual work on porting dialectal dictionaries and historical lexical resources developed at the Austrian Academy of Sciences onto representation languages that are supporting their publication in the Linked (Open) Data framework. We are aiming at a unified representation model that is flexible enough for describing those distinct types of lexical information. The goal is not only to be able to cross-link those resources, but also to link them in the Linked Data cloud with available data sets for highly-resourced languages and to elevate this way the dialectal and historical lexical resources to the same “digital dignity” as the mainstream languages have already gained.

12:30-13:00

Adding Dialectal Lexicalisations to Linked Open Data Resources: the Example of Alsatian

Delphine Bernhard

This article presents a method to align bilingual lexicons in a resource-poor dialect, namely Alsatian. One issue with Alsatian is that there is no standard and widely-acknowledged spelling convention and a lexeme may therefore have several different written variants. Our proposed method makes use of the double metaphone algorithm adapted to Alsatian in order to bridge the gap between different spellings. Once variant citation forms of the same lexeme have been aligned, they are mapped to BabelNet, a multilingual semantic network (Navigli and Ponzetto, 2012). The mapping relies on the French translations and on cognates for Alsatian words in the English and German languages.