Report for a CLARIN workshop type I

Towards Interoperability of Lexico-Semantic Resources

Date and location of the workshop

The workshop was held from January 31st to February 1st 2017 and was hosted by the Institute of Computer Science, University of Tartu, Estonia.

Information about the organizing team

Maciej Piasecki is an Associated Professor at Wrocław University of Science and Technology and a coordinator of the G4.19 Computational Linguistics and Language Technology Research Group. He holds PhD in Computer Science for work on Natural Language Processing. Maciej has been the leader of the Polish wordnet project since its beginning in 2005 till now. Maciej's main research topics are: extraction of the lexico-semantic knowledge from corpora, semi-automatic wordnet expansion, Distributional Semantics and relational lexical semantics. Maciej has been also working on morpho-syntactic processing of Polish (co-author of the first Polish tagger), Information Extraction, formal semantics and Machine Translation.

Erhard Hinrichs is a Full Professor (Ordinarius) for General and Computational Linguistics, Eberhard-Karls-Universität Tübingen, Seminar für Sprachwissenschaft, Tübingen, Germany, 1991 – present. He obtained Ph.D., Linguistics from The Ohio State University, Columbus, Ohio, USA. August 1985. He has been the leader of the GermaNet project since it beginning till now. Erhard Hinrichs is CLARIN National Coordinator of Germany.

Kadri Vider is a researcher of Language Technology at Institute of Computer Sciences in University of Tartu. She managed studies of Estonian Wordnet till 2007 when she started to work in Estonian Ministry of Education and Research arranging Estonian membership in CLARIN infrastructure also. Since coming back into University of Tartu in 2011 she manages Center of Estonian Language Resources - a research infrastructure consortium executing obligations of membership of Estonia in CLARIN ERIC. Kadri Vider is CLARIN National Coordinator of Estonia.

List of speakers and attendees

Invited speakers (3):

Francis Bond – Open Multilingual Wordnet, presented keynote speech ***Learning from each other: Linking and Enriching Wordnets***

Christiane Fellbaum – Princeton WordNet, presented keynote speech ***Lexical semantics for different resources: One WordNet does not fit all.***

German Rigau - Multilingual Central Repository, presented keynote speech ***Cross-lingual and Interoperable Event Mining in NewsReader***

Attendees (32):

Name	**CLARIN member state or organisation**	**Email**
Adam Rambousek	Czech Republic, Masaryk University	rambousek@fi.muni.cz
Ahti Lohk	Tallinn University of Technology	ahti.lohk@gmai.com
Andreas Witt	CLARIN-D	witt@ids-mannheim.de
António Branco	CLARIN Portuguese Language	Antonio.Branco@di.fc.ul.pt
Bolette Sandford Pedersen	UPCH, Denmark	bspedersen@hum.ku.dk
Christiane Fellbaum	Princeton University (not yet in CLARIN)	fellbaum@princeton.edu
Darja Fišer	Slovenia	darja.fiser@ff.uni-lj.si
Erhard Hinrichs	Tübingen University, CLARIN-D	erhard.hinrichs@uni-tuebingen.de
Ewa Rudnicka	CLARIN-PL	ewa.rudnicka@pwr.edu.pl
Fahad Khan	Italy - ILC-CNR	fahad.khan@ilc.cnr.it
Francis Bond	Nanyang Technological University	bond@ieee.org
German Rigau	University of the Basque Country	german.rigau@ehu.es
Heili Orav	University of Tartu	heili.orav@ut.ee
Helen Türk	University of Tartu	helen.turk@ut.ee
Indrek Jentson	CELR	indrek.jentson@ut.ee
John McCrae	National University of Ireland, Galway	john@mccr.ae
Kadri Vare	CELR	kadri.vare@ut.ee
Kadri Vider	EE, Center of Estonian Language Resources	kadri.vider@ut.ee
Kiril Simov	Bulgaria	kivs@bultreebank.org

Lars Borin	Sweden	lars.borin@svenska.gu.se
Lauma Pretkalniņa	Latvia / University of Latvia, Inst. of Mathematics and Computer Science	lauma.pretkalnina@gmail.com
Leo Võhandu	Tallinn University of Technology	lvohandu@gmail.com
Maciej Piasecki	CLARIN-PL	maciej.piasecki@pwr.edu.pl
Monica Monachini	Italy - ILC-CNR	monica.monachini@ilc.cnr.it
Neeme Kahusk	CELR	neeme.kahusk@ut.ee
Normunds Grūzītis	Latvia / University of Latvia, Inst. of Mathematics and Computer Science	normunds.gruzitis@lumii.lv
Peteris Paikens	Latvia / University of Latvia Institute of Mathematics and Computer Science	peteris@ailab.lv
Piotr Bański	CLARIN-D	banski@ids-mannheim.de
Sanni Nimb	Society for Danish Language & Literature	sn@dsl.dk
Sirli Zupping	University of Tartu	sirli.zupping@ut.ee
Tomasz Naskręt	CLARIN-PL	tomasz.naskret@pwr.edu.pl
Tõnis Nurk	Center of Estonian Language Resources	tonis.nurk@eki.ee

The goal of the workshop

The main goal of the workshop was to initiate the works on the improvement of interoperability, usability and ease of access of CLARIN L-SRs for (the needs of) their better visibility for H&SS users and their enhanced utilization in research applications. The key idea was to initiate coordinated development of a system of web services for accessing L-SRs and a common virtual Lexical Platform built on top of them. The platform is intended to be an open generic solution that will allow for effective linking, displaying and browsing of the rich variety of data included in CLARIN L-SRs. One of the functions of the platform will be a kind of federated search for L-SRs The platform will be an open system, implemented both as an open source code and open for all L-SRs. We can expect many potential installations and many web applications based on them.
The main topics discussed during the workshop included:

A. **Models of the lexico-semantic resources (L-SRs) from the interoperability perspective**
B. **Formats for the interoperability of lexico-semantic resources**
C. **User requirements and applications of the lexico-semantic resources**
D. **Interoperable platform for lexico-semantic resources: components, Web Services and integrators**

One of the outcomes of the workshop is an integrated report with an action plan for

each of the key topics discussed in the working groups. The report should be next offered to a wider CLARIN community for further discussion, testing and refinement. A limited set of standards and requirements should be identified for further work. After the workshop a joint proposal of CLARIN type II workshop will be formulated. This next workshop should be devoted to the construction of a first set of web services for accessing selected LRSs and the first prototype of the platform. The platform development is planned to be a collaborative work.

Contributions of the workshop to strategic goals of CLARIN ERIC

The topics of the workshop were closely related to several CLARIN strategic pillars (cf. CE-2015-0657), namely:

- *integration of data* – the common format and strategy of integration will form the basis for the interoperable multifaceted resources and will thus enable "the development of tools that allow for mapping between different standards"
- *integration of services* – the common platform and system of web services will simplify providing multilingual web services for semantic annotation, exploration, exploitation, enhancement, analysis, manipulation and visualization of language data; support for mono- and multilingual processing workflows,
- *ease of access* – the common platform and system of web services will simplify access to different types of lexico-semantic resources. They will will enhance multilingual browsing and visualization services.
- *crossing borders* – the common format, platform and strategy of integration will facilitate crossing borders between languages, countries, and infrastructures; "international collaboration both at the RI and at the thematic level".

Main results of the workshop

The workshop was divided into three main discussion sessions completed with a final planning session during which we summed up all the agreements made so far and formulated a plan of future activities.

First, different models of lexico-semantic resources (L-SRs) were discussed. We made a brief overview of basic notions and their interpretation across different types of L-SRs looking for some common points to facilitate interoperability. Most attention was given to the problem of identifying a common set of lexico-semantic relations that could be used in an common interlingual format or a pivotal semantic resource.

Second, the discussion entered a vast and problematic area of the existing formats for the interoperability of lexico-semantic resources. The general aim was to identify a set of formats used in the L-SRs represented during the workshop and next identify a limited subset of those formats that can be possibly generalised to the representation of most if not all L-SRs. RDF-based, LMF-based and Lemon-based formats were considered.

It seems that the further expansion of Lemon-based formats has significantly more advantages than the further development of any LMF-based variety of formats.

Next, we compiled a broad list of known and possible applications of L-SRs (mainly wordnets) encompassing the use of L-SRs in complex text mining systems and good practices in utilising L-SRs in research applications in Humanities and Social Sciences. The latter seem to be more limited than it could be expected. We have been also discussing ways of measuring the usage of resources including the problem of measuring the access statistics to individual resources in systems combining many L-SRs. The problem of L-SRs licenses was also discussed with the strong emphasis put on the promotion of open licenses. During the last, fourth session, an idea of an open virtual interoperable lexical platform was proposed. The platform is intended to link diversified lexical resources in a flexible and autonomous way. Each resource will be represented by a module with a predefined set of programming interfaces implemented as Web Services. The platform will be focused on browsing the resources by users as the primary task. The modules will be expected to return for the named resource elements their representations in an agreed presentation format.

As a result of the discussion, four main tasks were decided to be performed after the workshop: the definition of a common set of lexico-semantic relations, the development of a draft version of an extendable common format for wordnets, the formulation of a limited set of CMDI profiles for L-SRs and the design of the first version of the Lexical Platform associated with the implementation of the first prototype as a proof of concept.

Any publications planned

1. CAC'2017:
    a. L-SRs in CLARIN, problems with their interoperability and prospects,
    b. or design of the Lexical Platform and the first prototype, a more advanced version can be a topic of the potential post-CAC publication,
    c. CMDI profiles for L-SRs in VLO: problems and guidelines or suggested profiles.
2. Common relations: candidates common relations, as well as a first set of proposed common relations together with envisaged applications will be the topic of a paper submitted no later than for the Global WordNet Conference 2018 (GWC'18), the deadline in Sep. 2017 (LREC 2018 will be also considered)
3. Common format for wordnets that can be expanded and adapted to a wide range of wordnets will be described in a paper submitted no later than for GWC'18 (and/or LREC'18).

Potential impact and the next steps

The workshop was an important step towards the integration of groups developing L-SRs and further joint work on improving the visibility and interoperability of L-SRs in CLARIN. Even the brief overview showed the richness of resources offered to users, but not visible well enough. We have identified several problems:

1. diversified CMDI metadata assigned to L-SRs in CLARIN repositories (worsened

by potential technical problems in importing metadata),

2. lack of common basic notions in the description of the content of resources, especially a common list of the lexico-semantic relations,
3. lack of a common format for wordnets (the most numerous group of L-SRs), making construction of applications more difficult,
4. lack of one joint CLARIN searching and browsing facility across L-SRs, similar to Federated Content Search for corpora.

For all four problem areas working groups (task forces) were formed by volunteers. Even partial solving of the problems would contribute to the strategic goals targeted by the workshop. In task 4 we plan to work on a kind of Lexical Platform which can be a virtual aggregation of many L-SRs. It can become a central search tool offering functionality similar to FDC, but in the area of L-SRs. Finally, we discussed the preparation of a joint proposal for CLARIN type II workshop devoted to the construction of a first set of web services for accessing selected LRSs and a first prototype of the platform.

References

Annex 1: Discussion Overview and Tasks
https://docs.google.com/document/d/1RTx6m53vjdPuFRt-Mz3KqUJDUcd51LzK2kqJUNkipJg/edit?usp=sharing

Annex 2: Program with extended topics and links to presentations
https://docs.google.com/document/d/1X7nNhw9jLto5XaYIUP2uUZwQ6drVTTsvx8N9FfKC6iE/edit?usp=sharing