Working Group for Open Data in Linguistics: Status Quo and Perspectives

Christian Chiarcos+, Sebastian Hellmann*

+ Department of Linguistics, Collaborative Research Center (SFB) 632 "Information Structure", University of Potsdam
Karl-Liebknecht-Str. 25, 14476 Golm (Potsdam)
chiarcos@uni-potsdam.de
http://www.sfb632.uni-potsdam.de
* Universität Leipzig, Institut für Informatik, Johannisgasse 26,
D-04103 Leipzig, Germany
hellmann@informatik.uni-leipzig.de
http://aksw.org

Abstract. Since its formation in 2010, the Open Linguistics Working Group (OWLG) has been steadily growing and the direction the working group is heading has been clarified (although a number of issues remain open). We concentrated on the identification of goals and directions to pursue, and in this paper, we summarize results of this process, and talk about the current status of this working group as well as the main challenges and problems, we have identified so far.

1 Discussion results

An important result of our discussion are the seven points described in the next section, which define the purpose of the working group. In the next section, we summarize four major problems and challenges of the work with linguistic data. Such problems will become a primary topic of the Working Group. Thereafter, we give an overview of the current status and activities of the group and provide some suggestions for how to get involved.

1.1 Purpose

As a result of numerous discussions with interested linguists, NLP engineers and information technology experts, we identified seven open problems for our respective communities and their ways to use, to access and to share linguistic data. These represent the challenges to be addresses by the working group, and the role that it is going to fulfil:

1. Promote the idea and definition, as specified in opendefinition.org of open data in linguistics and in relation to language data.
2. Act as a central point of reference and support for people interested in open linguistic data.

3. Provide guidance on legal issues surrounding linguistic data to the community.
4. Build an index of indexes of open linguistic data sources and tools and link existing resources.
5. Facilitate communication between existing groups.
6. Serve as a mediator between providers and users of of technical infrastructure.
7. Assemble best-practice guidelines and use cases to create, use and distribute data.

In many aspects, the OWLG is not unique with respect to these goals. Indeed, there are numerous initiatives with similar motivation, e.g., the Cyberling blog[1], the ACL Special Interest Group for Annotation[2], and large multinational initiatives as the ISO initiative on Language Resources Management (ISO TC37/SC4)[3] or European projects such as CLARIN[4], FLARENET[5] and METANET[6]. The key difference between these and our Working Group is that we are not affiliated to an existing organization or one particular community, but that our members represent the whole band-width from academic linguistics (with its various subfields, e.g., typology and corpus linguistics) over applied linguistics (e.g., language documentation, computational linguistics, computational lexicography) and computational philology to natural language processing and information technology. We do not consider ourselves as being in competition with any existing organization, but hope to establish new links and further synergies between these. In the following section, we summarize typical and concrete scenarios where such an interdisciplinary community may help to resolve problems observed (or, sometimes, overlooked) in the daily praxis of working with linguistic resources.

2 Open linguistics resources, problems and challenges

Among the broad range of problems associated with linguistic resources, we identified four major classes of problems and challenges during our discussions that may be addressed by the OWLG. First, there is a great uncertainty with respect to **legal questions** of the creation and distribution of linguistic data; second, there are **technical problems** such as the choice of tools, representation formats and metadata standards for different types of linguistic annotation; third, we have not yet identified a point of reference for **existing open linguistic resources**; finally, there is the **agitation challenge**, i.e., how (and whether) we should convince our collaborators to release their data under open licenses. These challenges are described below in detail.

[1] http://cyberling.org/
[2] http://www.cs.vassar.edu/sigann/
[3] http://www.tc37sc4.org
[4] http://www.clarin.eu
[5] http://www.flarenet.eu
[6] http://www.meta-net.eu

Challenge 1: Legal questions

The linguistic community becomes increasingly aware of the potentially difficult legal status of different types of linguistic resources:

- How to find a suitable license for my corpus?
- Whose copyright do I have to respect? For example, corpora may have complex copyright situations where the original authors own the primary data, and thus may have partial copyright on the entire collection.
- Are there exceptions (e.g. for academic research) to the copyright that may allow me to work with my corpus anyway?
- How to circumvent (or solve) copyright issues?
- What legal restrictions apply to a particular resource (e.g., web corpora, newspaper corpora, digitizations of printed editions, audio and video files)?
- How to create multi-media (audio, video) data collections in a way that allows us to use (and hopefully, distribute) them for research?

The situation is even more complex because the legal situation may change over time (e.g., German copyright law was changed twice within the last decade), and this complexity multiplies on an international scale. The OWLG provides a platform to discuss such problems, to collect recommendations and document use cases as found in publications and technical reports, and discussed on conferences and mailing lists.

Challenge 2: Technical problems

Often, when creating a new corpus in a novel domain, the question is to be answered which tool to choose for which type of annotation. The OWLG will collect case studies and best practice recommendations with respect to this, it will encourage the documentation of use cases, collect links to documented case studies and best practice recommendations (e.g., by EMELD[7], or FLARENET[8]), and participate in the maintenance of existing sites that provide an overview over annotation tools and their domain of application (e.g., the Linguistic Annotation Wiki[9], or corresponding parts of the ACL Wiki[10]). A question related to the choice of tools is the question which representation formalisms to choose. We intend to provide basic information about proposed standard formats (e.g., the ISO proposal LAF/GrAF[11], the specifications of the Text Encoding Initiative [TEI][12]) and applicable formalisms (e.g., XML[13] or RDF[14]). These formats, again, are closely related to the question which corpus infrastructure (data

[7] http://emeld.org/school

[8] http://www.flarenet.eu/?q=Standards_and_Best_Practices

[9] http://annotation.exmaralda.org/index.php/Linguistic_Annotation

[10] http://aclweb.org/aclwiki/index.php?title=Tools_and_Software_for_English

[11] http://www.tc37sc4.org

[12] http://www.tei-c.org

[13] http://www.w3.org/XML

[14] http://www.w3.org/RDF

base, search interface) may be suitable to store, query and visualize what kind of linguistic annotations (e.g., domain- and community-specific tools like Toolbox[15] and ELAN[16], or general-purpose corpus query tools like ANNIS[17]). A third problem is the question of documentation requirements for different types of resources, the use of metadata standards (e.g., Dublin Core[18], or the TEI header[19]), and how annotation documentation and interoperability can be improved linking linguistic resources with terminology repositories (e.g., GOLD[20], ISOcat[21]). The OWLG aims to collect such questions and (partial) answers to these, we will contribute to existing metadata repositories and co-operate with other initiatives that pursue similar goals, e.g., the ACL Special Interest Group in Linguistic Annotation[22]. As opposed to these, the OWLG does not require membership in a particular organization, and we carry a focus on linguistic resources released under an open license. Further, we encourage (but do not require) the conversion of linguistic resources to Linked Data[23].

Challenge 3: Overview over existing resources

If a new research question is to be addressed, the question arises which resources may already be available and whether these may be accessible, and often, this problem is still solved by asking experts on mailing lists, e.g. the CORPORA list[24]. Therefore, the OWLG has begun to collect metadata about open linguistic resources within the CKAN repository[25]. Although there are other metadata repositories (e.g., those maintained by META-NET[26], FLARENET[27], or CLARIN[28]) available, the CKAN repository is qualitatively different in two respects: On the one hand, CKAN focuses on the license status of the resources and it encourages the use of open licenses. On the other hand, it is not specifically directed to linguistic resources, but rather, it is used by a large set of different working groups, whose resources may be exploited by linguists (e.g., exhaustive collections of legal documents from several countries [from law], or the open richly annotated cuneiform corpus [from archeology]).

[15] http://www.sil.org/computing/toolbox
[16] http://www.lat-mpi.eu/tools/elan
[17] http://www.sfb632.uni-potsdam.de/d1/annis
[18] http://dublincore.org
[19] http://www.tei-c.org/release/doc/tei-p5-doc/en/html/HD.html
[20] http://linguistics-ontology.org
[21] http://www.isocat.org
[22] http://www.cs.vassar.edu/sigann
[23] http://linkeddata.org
[24] http://listserv.linguistlist.org/archives/corpora.html
[25] http://ckan.net
[26] http://www.meta-net.eu
[27] http://www.flarenet.eu/?q=Documentation_about_Individual_Resources
[28] http://catalog.clarin.eu/ds/vlo

Challenge 4: Agitation

One of the goals of the OWLG is the promotion of open licenses for linguistic data collections. As we know from practical experience, researchers sometimes hesitate to provide their data under an open license. There has many different reasons for this, ranging from the uncertainty with respect to the legal situation to the (understandable) because fear that people exploit the resources before the original author had the chance to do so. We hope to contribute to the clarification of legal issues and to provide case studies that may help to clarify these problems. For example, one solution for second aspect mentioned above may be that data collections are designed as open linguistic resources from the beginning, but that their publication is delayed for several years, so that the creators can exploit this data long enough before any concurrent may get hands on it. One important argument that favors the use of open resources in academia is that only resources that are available to other researchers make it possible that empirically working linguists meet elementary scientific standards such as verifiability. Following this premise, we intend to promote the use of open resources in linguistics.

3 Current status and on-going developments (as of May, 19th, 2011)

So far, we focused on the task to delineate what questions the Open Linguistics Working Group may address, to formulate its general goals and potentially fruitful application scenarios. This paper summarizes these discussions, and it concludes a critical step in the formation process of the working group: Having defined a (preliminary) set of goals and principles, we can now concentrate on the tasks at hand, and in to collect resources and to attract interested people in order to address the challenges identified above. At the moment, our Working Group assembles 32 people from 21 different organizations and 7 countries (Germany, US, UK, France, Canada, Hungary, and Slovenia). Our group is relatively small, but continuously growing and sufficiently heterogeneous. It includes people from library science, typology, historical linguistics, cognitive science, computational linguistics, and information technology, just to name a few, so, the ground for fruitful interdisciplinary discussions has been laid out. We are very glad that famous linguists such as Nancy Ide (Text Encoding Initiative, American National Corpus, Vassar College) and Christiane Fellbaum (WordNet, University of Princeton) accepted our invitation to post guest blogs[29], and we would like to intensify this tradition and encourage all members of the OWLG to describe interesting projects and experiences on this medium, to share insights and difficulties over the Open Linguistics mailing list[30], and, of course, to join our meetings and telcos. The next meeting[31] is about to be held in conjunction with the Fifth Open Knowledge Conference (OKCon)[32], June 30th to July 1st 2011

[29] http://blog.okfn.org/taxonomy/working-groups/wg-linguistics/

[30] http://lists.okfn.org/mailman/listinfo/open-linguistics

[31] http://okcon.org/2011/programme/open-linguistics-workshop

[32] http://okfn.org/okcon/

in Berlin, Germany, and of course the OKCon itself is a great reason to join us there.[33]

As for our first concrete activities, we have begun to compile a **list of resources** of particular interest to the members of the working group. Most of these resources are free, others are partially free (i.e., annotations free, but text under copyright), and a few have been included that are very representative for a particular type of resource (e.g., corpora derived from the Penn Treebank as a prototypical multi-layer corpus). Altogether, the list comprises 102 entries by now, and the next step would be to register them at the CKAN metadata repository and to select a few for deeper investigation.

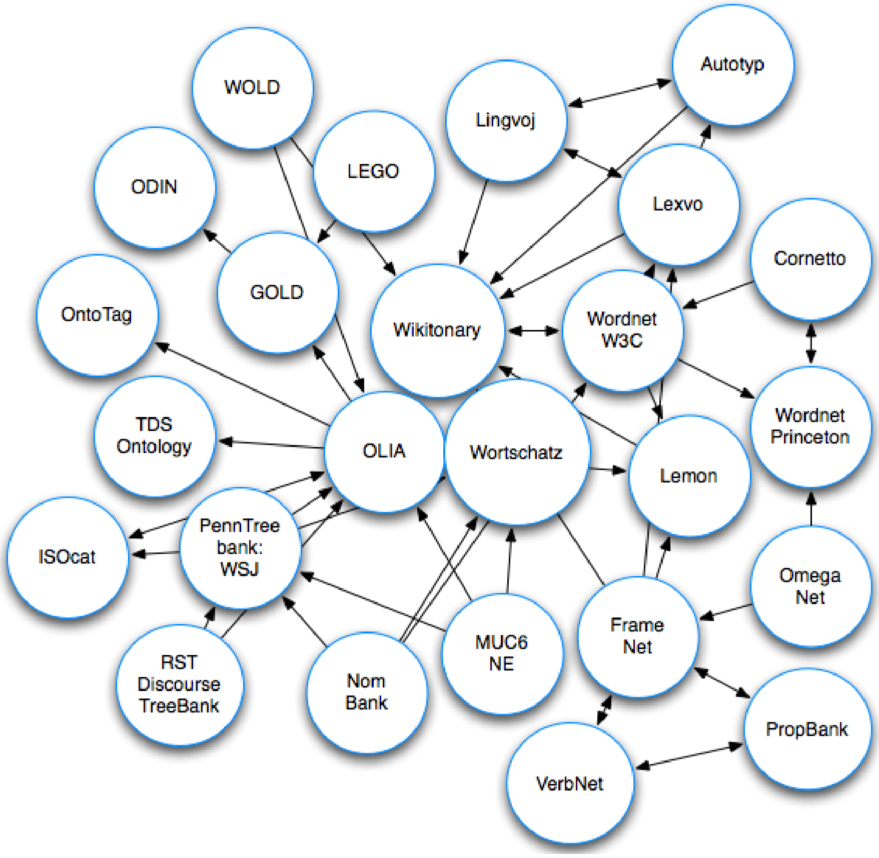

Fig. 1. Drafted version of the Linguistic LOD cloud. The official version by Cyganiak and Jentzsch can be found at `http://lod-cloud.net`.

[33] more ways to get involved can be found here: `http://linguistics.okfn.org/get-involved/`

One aspect of such investigations may be the conversion of some of the resources to RDF and to provide them as Linked Data. Several working group members (including the authors) are working towards this direction. The ultimate result may be an **Linguistics Linked (Open) Data** cloud, as sketched in Figure 1[34] . On this basis, novel applications in all participating fields may be developed.

[34] We would like to thank Amrapali Zaveri for custom-tailoring this figure for us.