The Open Linguistics Working Group of the Open Knowledge Foundation

Christian Chiarcos, Sebastian Hellmann, and Sebastian Nordhoff

Abstract The Open Linguistics Working Group (OWLG) is an initiative of experts from different fields concerned with linguistic data, including academic linguistics (e.g. typology, corpus linguistics), applied linguistics (e.g. computational linguistics, lexicography and language documentation) and NLP (e.g. from the Semantic Web community). The primary goals of the working group are 1) the promotion of the idea of open linguistic resources 2) the development of means for their representation, and 3) encouraging the exchange of ideas across different disciplines.
To a certain extent, the activities of the Open Linguistics Working Group converge towards the creation of a Linguistic Linked Open Data cloud, which is a topic addressed from different angles by several members. Some of these activities are described further in the other contributions of this part.

1 The Open Knowledge Foundation

The Open Knowledge Foundation (OKFN) is a not-for-profit organization founded in 2004 for the promotion of open knowledge, i.e. any kind of data and content that can be freely used, reused, and redistributed. Activities of the OKFN include the

Christian Chiarcos
Information Sciences Institute, University of Southern California, 4676 Admiralty Way # 1001, Marina del Rey, CA 90292 e-mail: chiarcos@daad-alumni.de

Sebastian Hellmann
Universität Leipzig, Fakultät für Mathematik und Informatik, Abt. Betriebliche Informationssysteme, Johannisgasse 26, 04103 Leipzig, Germany e-mail: hellmann@informatik.uni-leipzig.de

Sebastian Nordhoff
Department of Linguistics, Max Planck Institute for Evolutionary Anthropology, Deutscher Platz 6, 04103 Leipzig, Germany e-mail: sebastian_nordhoff@eva.mpg.de

development of standards (Open Definition), tools (CKAN) and support for working groups and events:

The **Open Definition** sets out principles to define 'openness' in relation to content and data.[1] This definition can be summed up in the statement that "A piece of content or data is open if anyone is free to use, reuse, and redistribute it – subject only, at most, to the requirement to attribute and share-alike."

The OKFN provides a catalog system for open datasets, **CKAN**.[2] CKAN is an open-source data portal software developed to publish, to find, and to reuse open content and data easily, especially in ways that are machine automatable.

The OKFN also serves as host for various working groups addressing problems of open data in different domains. At the time of writing, there are 18 OKFN **working groups** covering fields ranging from government data and economics over archeology or open text books to cultural heritage.[3] The OKFN organizes various events such as the Open Knowledge Conference (OKCon) and facilitates the communication between different working groups.

In 2010, the **OKFN Working Group on Open Linguistic Data** (OWLG) was founded. Since its formation, the Open Linguistics Working Group has been steadily growing. We have identified goals and problems that are to be addressed, and directions that are to be pursued in the future. Preliminary results of this on-going discussion process are summarized in this contribution: Section 2 specifies the goals of the working group by identifying seven points that emerged from our discussions; Sect. 3 identifies four major problems and challenges of the work with linguistic data; and, finally, Sect. 4 gives an overview of recent activities and the current status of the group.

2 Goals of the Open Linguistics Working Group

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 fulfill:

1. Promote the idea 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.

[1] http://www.opendefinition.org

[2] http://ckan.org/

[3] For a complete overview see http://okfn.org/wg.

6. Serve as a mediator between providers and users 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 and overlapping goals, e.g. the Cyberling blog,[4] the ACL Special Interest Group for Annotation (SIGANN),[5] and large multi-national initiatives such as the ISO initiative on Language Resources Management (ISO TC37/SC4),[6] the American initiative on Sustainable Interoperability of Language Technology (SILT),[7] or European projects such as the initiative on Common Language Resources and Technology Infrastructure (CLARIN),[8] the Fostering Language Resources Network (FLaReNet),[9] and the Multilingual Europe Technology Alliance (META).[10]

The key difference between these and the OWLG is that we are not grounded within a *single* community, or even restricted to a hand-picked set of collaborating partners, 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 to be in competition with any existing organization or initiative, but we hope to establish new links and further synergies between these. Section 3 summarizes typical and concrete scenarios where such an interdisciplinary community may help to resolve problems observed (or, sometimes, overlooked) in the daily practice of working with linguistic resources.

3 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 that may be addressed by the OWLG:

legal questions There is a great uncertainty with respect to legal questions of the creation and distribution of linguistic data.

technical problems Often, researchers come up with questions regarding the choice of tools, representation formats and metadata standards for different types of linguistic annotation.

[4] http://cyberling.org/

[5] http://www.cs.vassar.edu/sigann/

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

[7] http://www.anc.org/SILT

[8] http://www.clarin.eu

[9] http://www.flarenet.eu

[10] http://www.meta-net.eu

repository of open linguistic resources So far, the communities involved have not yet established a common point of reference for existing open linguistic resources; at the moment there are multiple metadata collections.

spread the word Finally, there is an agitation challenge for open data in linguistics, i.e. how (and whether) we should convince our collaborators to release their data under open licenses.

3.1 Legal Questions

The linguistic community is increasingly becoming aware of the potentially difficult legal status of different types of linguistic resources. Typical questions include: *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) ?*, or *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, and this complexity multiplies on an international scale. The OWLG can provide 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.

3.2 Technical Problems

When creating a new corpus in a novel domain, researchers are confronted with the question which tool to choose for which type of annotation. The OWLG can 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 the American project on Electronic Metadata for Endangered Languages Data (EMELD, 2002-2007),[11] or FLaReNet[12]), and participate in the maintenance of existing sites that provide an overview of annota-

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

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

tion tools and their domains of application (e.g. the Linguistic Annotation Wiki,[13] or corresponding parts of the Wiki of the Association of Computational Linguistics[14]).

A question related to the choice of tools is the choice of representation formalisms. Interoperability between different representation formalisms used in NLP pipelines, linguistic corpora and other types of linguistic resources has become a major field of research in the last years. We intend to provide basic information about proposed standard formats (e.g. the ISO TC37/SC4 proposal LAF/GrAF (Ide and Suderman, 2007)), the specifications of the Text Encoding Initiative (Ide and Veronis, 1995, TEI)) and applicable formalisms (e.g. XML or RDF). In this volume, several publications addressed interoperability issues and possible solutions based on RDF and related formalisms.

These formats, again, are closely related to the question which corpus infrastructure (data 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 (Busemann and Busemann, 2008) and ELAN (Hellwig et al, 2008), or general-purpose corpus query tools like ANNIS (Zeldes et al, 2009)). A third problem is the question of documentation requirements for different types of resources, the use of metadata standards (e.g. Dublin Core (Weibel et al, 1998), or the TEI header (Giordano, 1995), and how annotation documentation and interoperability can be improved by linking linguistic resources with terminology repositories (e.g. the General Ontology of Linguistic Description GOLD (Farrar and Langendoen, 2003), or the ISO TC37/SC4 Data Category Registry (Kemps-Snijders et al, 2009). 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 (SIGANN).[15]

Within the working group, we encourage (but do not require) the conversion of linguistic resources to Linked Data,[16] and selected activities in this direction currently conducted by different OWLG members are described here by Chiarcos (this vol.), Hellmann et al (this vol.), Nordhoff (this vol.) and Chiarcos et al (this vol.) who discuss formalisms and technologies for the modeling and the interlinking of representative types of linguistic data in a Linguistic Linked Open Data cloud.

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

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

[15] http://www.cs.vassar.edu/sigann

[16] http://linkeddata.org

3.3 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. Often, this problem is still solved by asking experts on mailing lists.[17]

In order to establish a more structured way of collecting information, the OWLG has begun to collect metadata about open linguistic resources within the CKAN repository.[18] CKAN is qualitatively different from earlier metadata repositories[19] in two respects:

- CKAN focuses on the license status of the resources and it encourages the use of **open** licenses.
- CKAN is **not restricted to linguistic resources**, but rather, it is used by all working groups, as well as interested individuals outside these working groups. Examples for CKAN resources created outside the linguistic/NLP community that are nevertheless of relevance to linguists include collections of open textbooks,[20] the complete works of Shakespeare,[21] or the Open Richly Annotated Cuneiform Corpus (ORACC).[22]

3.4 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. This is partially due to the uncertainty with respect to the legal situation, but there are also certain **sociological** factors, e.g. the (understandable) fear that people exploit the resources before the original author had the chance to do so.

We hope to encourage the discussion of legal issues and to provide case studies that may help clarify these problems. For example, a solution for the sociological 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 make private use of their 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

[17] For example, the CORPORA mailing list, `http://listserv.linguistlist.org/archives/corpora.html`

[18] `http://ckan.net`, resources tagged by `linguistics`.

[19] For example, those maintained by META-NET (`http://www.meta-net.eu`) FLaReNet (`http://www.flarenet.eu/?q=Documentation_about_Individual_Resources`) or CLARIN (`http://catalog.clarin.eu/ds/vlo`.)

[20] `http://wiki.okfn.org/Wg/opentextbooks`

[21] `http://openshakespeare.org`

[22] `http://oracc.museum.upenn.edu`

make it possible that linguists working empirically meet elementary scientific standards such as verifiability.

4 Recent Activities and Ongoing Developments

In the first year of its existence, the OWLG focused on the questions to address, the formulation of general goals, and the identification of potentially fruitful application scenarios. At the time of writing, we have reached 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, e.g. to collect resources and to attract interested people in order to address the challenges identified above.

As of October, 12th, 2011, the Working Group assembles 67 people from 29 different organizations and 10 countries.[23] 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.

The Working Group maintains a home page,[24] a mailing list,[25] a wiki,[26] and a (guest) blog,[27] currently featuring contributions by Nancy Ide (Text Encoding Initiative, American National Corpus, Vassar College) and Christiane Fellbaum (WordNet, University of Princeton). We conduct regular meetings and have established a series of workshops (Workshop on Open Data in Linguistics at the 6th Open Knowledge Conference (OKCon 2011, Berlin, Germany, June 30th, 2011); Workshop on Linked Data in Linguistics (LDL 2012, Frankfurt/M., Germany, March 7th-9th)).

Recent community activities include the compilation of a list of resources that represent interesting candidates for the Linguistic Linked Open Data cloud. 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 not free, but very representative for a particular type of resource (e.g. corpora derived from the Penn Treebank (Marcus et al, 1994) as prototypical examples for multi-layer corpora). As of Dec, 22th, 2011, this list comprises 103 resources, including lexicons, word lists, corpora, and collections of linguistic metadata. Subsequently, these resources are registered at the CKAN metadata repository and a few will be selected for deeper investigation. The development of a collection of open, freely accessible linguistic resources that are represented in interoperable standards represents a concrete goal for several members of the working group and may be seen as a long-term vision of the OWLG.

[23] Germany, US, UK, France, Canada, Australia, the Netherlands, Greece, Hungary, Slovenia

[24] `http://linguistics.okfn.org`

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

[26] `http://wiki.okfn.org/Wg/linguistics`

[27] `http://blog.okfn.org/category/working-groups/wg-linguistics`

The identification of these resources will be followed by the analysis of possible links between them, and the generation of a Linguistic Linked Data Cloud. A summary of first steps into this direction are described by Chiarcos et al (this vol.).

References

Busemann A, Busemann K (2008) Toolbox self-training. Tech. rep., http://www.sil.org, version 1.5.4, Oct 2008

Chiarcos C (this vol.) Interoperability of corpora and annotations. P. 161-179

Chiarcos C, Hellmann S, Nordhoff S (this vol.) Linking linguistic resources: Examples from the Open Linguistics Working Group. P. 201-216

Farrar S, Langendoen DT (2003) A Linguistic Ontology for the Semantic Web. GLOT International 7:97–100

Giordano R (1995) The TEI header and the documentation of electronic texts. Computers and the Humanities 29(1):75–84

Hellmann S, Stadler C, Lehmann J (this vol.) The German DBpedia: A sense repository for linking entities. P. 181-189

Hellwig B, Uytvanck DV, Hulsbosch M (2008) ELAN - Linguistic Annotator. Tech. rep., http://www.lat-mpi.eu/tools/elan, version of 2008-07-31

Ide N, Suderman K (2007) GrAF: A graph-based format for linguistic annotations. In: Proc. Linguistic Annotation Workshop (LAW 2007), Prague, Czech Republic, pp 1–8

Ide N, Veronis J (1995) Text encoding initiative: Background and contexts. Kluwer Academic Pub

Kemps-Snijders M, Windhouwer M, Wittenburg P, Wright S (2009) ISOcat: Remodelling metadata for language resources. International Journal of Metadata, Semantics and Ontologies 4(4):261–276

Marcus M, Santorini B, Marcinkiewicz M (1994) Building a large annotated corpus of English: The Penn Treebank. Computational Linguistics 19(2):313–330

Nordhoff S (this vol.) Linked Data for linguistic diversity research: Glottolog/Langdoc and ASJP. P. 191-200

Weibel S, Kunze J, Lagoze C, , Wolf M (1998) RFC 2413 - Dublin Core metadata for resource discovery. http://www.isi.edu/in-notes/rfc2413.txt

Zeldes A, Ritz J, Lüdeling A, Chiarcos C (2009) ANNIS: A search tool for multilayer annotated corpora. In: Proc. Corpus Linguistics, Liverpool, UK, pp 20–23