Introduction and Overview

Christian Chiarcos, Sebastian Hellmann, and Sebastian Nordhoff

1 Linked Data in Linguistics

The explosion of information technology in the last two decades has led to a substantial growth in quantity, diversity and complexity of web-accessible linguistic data. These resources become even more useful when linked with each other, and the last few years have seen the emergence of numerous approaches in various disciplines concerned with linguistic resources.

It is the challenge of our time to store, interlink and exploit this wealth of data accumulated in more than half a century of computational linguistics (Dostert, 1955), of empirical, corpus-based study of language (Francis and Kucera, 1964), and of computational lexicography (Morris, 1969) in all its heterogeneity.

A crucial question involved here is the **interoperability** of the language resources, actively addressed by the community since the late 1980s (Text Encoding Initiative, 1990), but still a problem that is partially solved at best (Ide and Pustejovsky, 2010). A closely related challenge is **information integration**, i.e., how heterogeneous information from different sources can be retrieved and combined in an efficient way.

With the rise of the Semantic Web, new representation formalisms and novel technologies have become available, and, independently from each other, researchers in different communities have recognized the potential of these developments with

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

respect to the challenges posited by the heterogeneity and multitude of linguistic resources available today. Many of these approaches follow the **Linked Data paradigm** (Berners-Lee, 2006, see below) that postulates rules for the publication and representation of web resources. If (linguistic) resources are published in accordance with these rules, it is possible to follow links between existing resources to find other, related data and exploit network effects.

This volume provides an overview of the broad variety of approaches towards the application of the Linked Data paradigm to linguistic resources. It assembles the contributions of the workshop on Linked Data in Linguistics (LDL-2012), held at the 34th Annual Meeting of the German Linguistic Society (Deutsche Gesellschaft für Sprachwissenschaft, DGfS), March 7th-9th, 2012, in Frankfurt/M., Germany, organized by the Open Linguistics Working Group (OWLG)[1] of the Open Knowledge Foundation (OKFN),[2] an initiative of experts from different fields concerned with linguistic data, including academic linguists (e.g., typology, corpus linguistics), applied linguistics (e.g., computational linguistics, lexicography and language documentation), and NLP engineers (e.g., from the Semantic Web community). The primary goal of the working group is to promote the idea of open linguistic resources, to develop means for their representation, and to encourage the exchange of ideas across different disciplines. Accordingly, the current volume represents a great band-width of contributions from various fields, representing principles, use cases, and best practices for using the Linked Data paradigm to represent, exploit, store, and connect different types of linguistic data collections.

One goal of this book and the workshop on Linked Data in Linguistics (LDL-2012) is to document and to summarize these developments, and to serve as a point of orientation in the emerging domain of research on Linked Data in Linguistics. This documentary goal is complemented by social goals: (a) to facilitate the communication between researchers from different fields who work on linguistic data within the Linked Data paradigm; and (b) to explore possible synergies and to build bridges between the respective communities, ranging from academic research in the fields of language documentation, typology, translation studies, digital humanities in general, corpus linguistics, computational lexicography and computational linguistics, and computational lexicography to concrete applications in Information Technology, e.g., machine translation, or localization.

2 Technological Background

Several standards developed by different initiatives are referenced or used throughout this work. One is the **Extensible Markup Language** (XML, Bray et al, 1997) and its predecessor, the Standard Generalized Markup Language (SGML, Goldfarb and Rubinsky, 1990). These are text-based formats that allow to encode documents

[1] http://linguistics.okfn.org

[2] http://okfn.org

in an appropriate way for representing and transmitting machine-readable information.

XML and SGML have been the basis for most proposals for **interoperable representation formalisms specifically for linguistic resources**, for example the Corpus Encoding Standard (CES, Ide, 1998) developed by the Text Encoding Initiative (TEI[3]), or the Graph Annotation Format (GrAF, Ide and Suderman, 2007) developed in the context of the Linguistic Annotation Framework (LAF) by ISO TC37/SC4[4]. Earlier standards for linguistic corpora used XML data structures (i.e., trees) directly, but since Bird and Liberman (2001), it is generally accepted that generic formats to represent linguistic annotations should be based on graphs. State-of-the-art formalisms for linguistic corpora follow this assumption, and represent linguistic annotations in XML standoff formats, i.e., as bundles of XML files that are interlinked with cross-references, e.g., with formats like ATLAS (Bird and Liberman, 2001), PAULA XML (Dipper, 2005), or GrAF (Ide and Suderman, 2007).

In parallel to these formalisms, which are specific to linguistic resources, other communities have developed the **Resource Description Framework** (RDF, Lassila and Swick, 1999). Although RDF was originally invented to provide formal means to describe resources, e.g. books in a library or in an electronic archive (hence its name), its data structures were so general that its use has extended far beyond the original application scenario. RDF is based on the notion of **triples** (or 'statements'), consisting of a **predicate** that links a **subject** to an **object**. In other words, RDF formalizes relations between resources as labeled edges in a directed graph. Subjects are represented using globally unique Uniform Resource identifiers (URIs) and point (via the predicate) to another URI, the object part, to form a graph. (Alternatively, triples can have simple strings in the object part that annotate the subject resource.) At the moment, RDF represents the primary data structure of the Semantic Web, and is maintained by a comparably large and active community. Further, it provides crucial advantages for the publication of linguistic resources in particular: RDF provides a graph-based data model as required by state-of-the-art approaches on generic formats for linguistic corpora, and several RDF extensions were specifically designed with the goal to formalize knowledge bases like terminology data bases and lexical-semantic resources. For resources published under open licenses, an RDF representation yields the additional advantage that resources can be interlinked, and it is to be expected that an additional gain of information arises from the resulting network of resources. If modeled with RDF, linguistic resources are thus not only **structurally interoperable** (using RDF as representation formalism), but also **conceptually interoperable** (with metadata and annotations are modeled in RDF, different resources can be directly linked to a single repository). Further, concrete applications using linguistic resources can be build on the basis of the rich ecosystem of format extensions and technologies that has evolved around RDF, including APIs, RDF databases (triple stores), the query language SPARQL, data browsing and visualization tools, etc.

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

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

For the formalization of knowledge bases, several RDF extensions have been provided, for example the **Simple Knowledge Organization System** (SKOS, Miles and Bechhofer, 2009), which is naturally applicable to lexical-semantic resources, e.g., thesauri. A thorough logical modeling can be achieved by formalizing linguistic resources as ontologies, using the **Web Ontology Language** (OWL, McGuinness and Van Harmelen, 2004), another RDF extension. OWL comes in several dialects (profiles), the most important being OWL/DL and its sublanguages (e.g. OWL/Lite, OWL/EL, etc.) that have been designed to balance expressiveness and reasoning complexity (McGuinness and Van Harmelen, 2004; W3C OWL Working Group, 2009) OWL/DL is based on Description Logics (DL, Baader et al, 2005) and thus corresponds to a *decidable* fragment of first-order predicate logic. A number of reasoners exist that can draw inferences from an OWL/DL ontology and verify consistency constraints. Primary entities of OWL Ontologies are **concepts** that correspond to classes of objects, **individuals** that represent instances of these concepts, and **properties** that describe relations between individuals. Ontologies further support **class operators** (e.g. intersection, join, complement, instanceOf, subClassOf), as well as the specification of **axioms** that constrain the relations between individuals, properties and classes (e.g. for property *P*, an individual of class *A* may only be assigned an individual of class *B*). As OWL is an extension of RDF, every OWL construct can be represented as a set of RDF triples.

RDF is based on globally unique and accessible URIs and it was specifically designed to establish links between such URIs (or resources). This is captured in the **Linked Data paradigm** Berners-Lee (2006) that postulates four rules:

1. Referred entities should be designated by URIs,
2. these URIs should be resolvable over HTTP,
3. data should be represented by means of standards such as RDF,
4. and a resource should include links to other resources.

With these rules, it is possible to follow links between existing resources to find other, related, data and exploit network effects. The **Linked Open Data (LOD) cloud**[5] represents the resulting set of resources. If published as Linked Data, linguistic resources represented in RDF can be linked with resources already available in the Linked Open Data cloud. At the moment, the LOD cloud covers a number of lexico-semantic resources, including the Open Data Thesaurus,[6] WordNet,[7] Cornetto (Dutch WordNet),[8] DBpedia (machine-readable version of the Wikipedia),[9] Freebase (an entity database),[10] OpenCyc (database of real-world con-

[5] http://lod-cloud.net

[6] http://vocabulary.semantic-web.at/PoolParty/wiki/OpenData

[7] http://semanticweb.cs.vu.nl/lod/wn30, http://www.w3.org/TR/wordnet-rdf, http://wordnet.rkbexplorer.com

[8] http://www2.let.vu.nl/oz/cltl/cornetto

[9] http://www.dbpedia.org

[10] http://freebase.com

cepts),[11] and YAGO (a semantic knowledge base).[12] Additionally, the LOD cloud includes knowledge bases of information about languages and bibliographical information that are relevant for here, e.g., Lexvo (metadata about languages),[13] lingvoj (metadata about language in general),[14] Project Gutenberg (bibliographical data base)[15] and the OpenLibrary (bibliographical data base).[16] Given the interest that researchers take in representing linguistic resources as Linked Data, continuing growth of this set of resources seems to be assured. Several contributions assembled in this volume discuss the linking of their resources with the Linked Open Data cloud, thereby supporting the overarching vision of a Linguistic Open Data (sub-) cloud of linguistic resources, a **Linguistic Linked Open Data cloud** (LLOD).

The workshop on Linked Data in Linguistics (LDL-2012) was the first major event organized by the Open Linguistics Working Group (OWLG), and we hope that the workshop and this volume contribute to the on-going formation of an interdisciplinary community actively working towards the application of the Linked Open Data paradigm to all forms of linguistic resources, that they facilitate the exchange of ideas and resources with the long-term goal to build such a Linguistic Linked Open Data cloud.

One goal of this book is to document and to summarize recent developments in this direction, and to serve as a point of orientation to where research on Linked Data in Linguistics is heading to. Almost more important, however, is the second goal we pursued, i.e., to facilitate the communication between researchers working in this direction, to explore possible synergies and to build bridges between these communities, and we would like to thank the participants of the Linked Data in Linguistics workshop as well as the members of the OWLG for sharing their ideas, insights and/or resources, and we hope that, together, we can build a Linked Data (sub)cloud of linguistic resources that can be used across different disciplines for mutual benefit of researchers and the developers of NLP and Semantic Web applications.

2.1 Notational Conventions

Throughout this volume, the following notational conventions are applied:

- linguistic examples are written in a *slanted font*, translations are marked with single quotes
- ontological concepts, source code, URLs and URIs are written in a `typewriter font`

[11] http://sw.opencyc.org

[12] http://mpii.de/yago

[13] http://www.lexvo.org

[14] http://www.lingvoj.org

[15] http://www4.wiwiss.fu-berlin.de/gutendata

[16] http://openlibrary.org

- RDF literals are marked by single or double quotes

3 Structure of this Volume

We are happy to have attracted a large number of high quality contributions from very different domains for the workshop on Linked Data in Linguistics (LDL-2012) held March 7^{th} - 9^{th}, 2012, as part of the 34^{th} Annual Meeting of the German Linguistics Society (DGfS) in Frankfurt a. M., Germany. The set of subdisciplines included in this volume is diverse; the goal is the same: provide scientific data in an open format which permits integration with other data repositories.

This volume is organized in four parts: Parts I, II and III describe applications of the Linked Data paradigm to major types of linguistic resources, i.e., **lexical-semantic resources**, **linguistic corpora** and **other knowledge bases**, respectively. These parts represent the contributions of the participants of the Workshop Linked Data in Linguistics (LDL-2012). In Part IV, the editors describe recent efforts to **link linguistic resources** – and thus to create a Linked Open Data (sub-)cloud of linguistic resources – in the context of the Open Linguistic Working Group (OWLG) of the Open Knowledge Foundation (OKFN). They illustrate how lexical-semantic resources, corpora and other linguistic knowledge bases can be interlinked and what possible gains of information are to be expected, using representative examples for the respective classes of linguistic resources.

As we are interested in linking different language resources, it should be noted that there is a natural overlap between these categories, and therefore, many contributions could be classified under more than one category. Bouda and Cysouw (this vol.), for example, discuss not only lexical resources, but also corpus representation, and knowledge bases for linguistic metadata; Schalley (this vol.) and Declerck et al (this vol.) describe not only linguistic knowledge bases, but also corpus data and multi-layer annotations; and the contributions by Chiarcos (this vol.), Hellmann et al (this vol.), and Nordhoff (this vol.) that are presented in the context of linking linguistic resources, could also have been presented in the respective parts on linguistic corpora, lexical-semantic resources and other (linguistic) knowledgebases.

3.1 Lexical Resources

Part I describes the modeling of various lexical-semantic resources as illustrated for lexical-semantic resources.

Peter Bouda and Michael Cysouw describe the digitization of dictionaries, and how the elements (head words, translations, annotations) found in there can be served in a Linked Data way while at the same time maintaining access to the document in its original form. To this end, they use standoff markup, which furthermore allows the third-party annotation of their data. They also explore how these third-

party annotations could be shared in novel ways beyond the normal scope of normal academic distribution channels, e.g. Twitter.

John McCrae, Elena Montiel-Ponsoda and Philipp Cimiano describe the *lemon* format that has been developed for the sharing of lexica and machine readable dictionaries. They consider two resources that seem ideal candidates for the Linked Data cloud, namely WordNet 3.0 and Wiktionary, a large document based dictionary. The authors discuss the challenges of converting both resources to *lemon*, and in particular for Wiktionary, the challenge of processing the mark-up, and handling inconsistencies and underspecification in the source material. Finally, they turn to the task of creating links between the two resources and present a novel algorithm for linking lexica as lexical Linked Data.

Axel Herold, Lothar Lemnitzer, and Alexander Geyken report on the lexical resources of the long-term project 'Digitales Wörterbuch der deutschen Sprache' (DWDS) which aims at the integration of several lexical and textual resources in order to document the German language and its use at several stages. They describe the explicit linking of four lexical resources on the level of individual articles which is achieved via a common meta-index. The authors present strategies for the actual dictionary alignment as well as a discussion of models that can adequately describe complex relations between entries of different dictionaries.

David Lewis et al. describe perspectives of Linked Data in the fields of software localisation and translation. They present a platform architecture for sharing, searching and interlinking of Linked Localisation and Language Data on the web. This architecture rests upon a semantic schema for the respective resources that is compatible with existing localisation data exchange standards and can be used to support the round-trip sharing of language resources. The paper describes the development of the schema and data management processes, web-based tools and data sharing infrastructure that use it. An initial proof of concept prototype is presented which implements a web application that segments and machine translates content for crowd-sourced post-editing and rating.

3.2 Linguistic Corpora

Part II deals with problems to create, to maintain and to evaluate linguistic corpora and other collections of linguistically annotated data. Previous research indicates that formalisms such as RDF and OWL are suitable to represent linguistic annotations Burchardt et al (2008); Cassidy (2010) and to build NLP architectures on this basis Wilcock (2007); Hellmann (2010), yet so far, it has rarely been applied to this type of linguistic resource.

Marieke van Erp describes interoperability problems of linguistic resources, in particular corpora, and develops a vision to apply the Linked Data approach to these issues. In her contribution, the constraints for linguistic resource reuse and the tasks are detailed, accompanied by a Linked Data approach to standardise and reconcile concepts and representations used in linguistic annotations.

As mentioned above, these problems are addressed in the NLP community by generic data models for linguistic corpora that are based on directed graphs. Kerstin Eckart, Arndt Riester and Katrin Schweitzer describe such a state-of-the-art approach on the task of resource integration for multiple independent layers of annotation in a multi-layer annotated corpus that is based on a graph-based data model, although not on RDF, but an XML standoff format and a relational database management system. They present an annotated corpus of German radio news including syntactic information from a parser, as well as manually annotated information status labels and prosodic labels. They describe each annotation layer and focus on the linking of the data from both layers of annotation, and show how the resource can support data extraction on both annotation layers. Although they do not directly make use of the Linked Data paradigm, the problems identified and the data model employed represent important steps towards the development of representation formalisms for multi-layer corpora by means of RDF and as Linked Data, see, for example, Chiarcos (this vol.).

Michael Carl and Henrik Høeg Müller describe a fascinating intersection between pure structural syntactic data and human-machine interaction in translation processes. Human behaviour while translating on a computer can be recorded with eye trackers and capturing of user input (mouse, keyboard). This behavioural data can then be linked to syntactic data extracted from the sentence translated (constituency, dependency). The intuition is that syntactically complicated sentences will have a repercussion in the user behaviour (longer gaze, slower input, more corrections). Carl and Müller, just like Bouda and Cysouw, and Eckart et al., use standoff annotation to allow for overlapping annotations. Their use of structural data on the one hand and behavioural data from a novel domain on the other hand shows the benefits the provision of data as Linked Data can have.

María Blume, Suzanne Flynn and Barbara Lust describe DTA, an online tool for the study of language acquisition. DTA allows for data creation, data management and collaborative use of child language data from a variety of languages (Spanish, French, English, Sinhala). Language Acquisition is a relative newcomer to the area of Linked Data, and it is exciting to see that areas somewhat distant from the NLP origins of Linked Data are beginning to join the movement.

3.3 Linguistic Knowledgebases

While Part II focused on annotated linguistic data, Part III presents a number of repositories of knowledge about languages and linguistic terminology that can be used, for example, for annotating linguistic data with linguistic analyses and metadata.

Menzo Windhouwer and Sue Ellen Wright describe the linking from language resources to linguistic data categories in ISOcat, a repository of linguistic terminology developed to foster semantic interoperability of linguistic resources. This registry follows a grass roots approach, which means that any linguist can add the

data categories (s)he needs. However, the goal of improving semantic interoperability can only be met if the data categories are reused by a wide variety of linguistic resource types. A resource indicates its usage of data categories by linking to them, this paper describes the technical prerequisites to achieve this in an RDF-based approach.

Thierry Declerck et al. describe strategies for exploiting the large set of dynamically increasing, freely available language data incorporated in the Linked Open Data (LOD) framework. Such language data currently mostly exist in the form of raw, unstructured textual expressions within RDF labels or comments. Incorporating them as structured language data within the LOD leads to a linguistic enrichment of the data sets that express linked (domain) knowledge resources, and this will enable the creation of more accurate, knowledge-aware NLP applications. This integration of linguistic information in knowledge representation systems should be done in compliance with both ISO (multi-layer linguistic annotation and data categories) and W3C (RDF, SKOS) standards. By this, new linguistically enriched datasets can also be more easily ported into the LOD format: e.g., repositories in the field of Digital Humanities often hold language data in taxonomical structures. The potential of linked language data for digital humanities is illustrated here for the detection of motifs in literary texts. For this purpose, a formal representation of the taxonomical structure of the Thompson Motif-Index of folk-literature (Thompson, 1955-58, TMI) is presented.

In a similar vein, Antonia Pareja-Lora reports on the development of a concept taxonomy for a different type of linguistic annotation, namely pragmatic annotations. Pragmatics has to deal with a real mix of different linguistic topics, such as (i) speech acts, (ii) deixis, presuppositions and implicatures; or (iii) pragmatic coherence relations, which traditionally have been tackled following several fragmentary and/or partial approaches. Pareja-Lora describes an approach to specify formally the different elements that a pragmatic annotation scheme should contemplate and make explicit with the goal to facilitate the interoperability of linguistic annotations up to the pragmatic level.

While the terminology repositories and taxonomies described in this part so far have been developed for interoperability of NLP tools and linguistic annotations, the remaining chapters of this part deal with typological databases that provide information about languages from a slightly different angle of research.

Steve Moran tackles the very basic unit of linguistics, the phoneme, and shows how heterogeneous data bases of phoneme inventories found in the worlds languages can be integrated with a Linked Data approach via mapping of the relations found in the original data bases to his ontology. His system is in production stage, and Moran shows how a number of phonological hypotheses can be confirmed or refuted using his PHOIBLE database. Moran furthermore explores the difference between queries in traditional relational databases and SPARQL queries.

Andrea Schalley casts a wide net and lists the criteria a typological knowledge base would have to respond to in an ideal world. She then discusses challenges for the realization and sketches the development of a computational tool that utilises Semantic Web technologies in order to provide novel ways to process, integrate, and

query cross-linguistic data. Its data store incorporates a set of ontologies (comprising linguistic examples, annotations, language background information, and metadata) backed by a software logic reasoner. This allows for highly targeted querying and answers on rather specific questions such as (i) which size (in terms of speaker count) do languages have that have kin-sensitive pronouns?, or (ii) which languages code joint attention in their grammar, and if so, where in the grammar do they do it?

3.4 Towards a Linguistic Linked Open Data Cloud

The last part describes joint activities of different members of the Open Linguistics Working Group (OWLG) aiming to develop a Linked Open Data (sub-)cloud of linguistic resources.

Christian Chiarcos, Sebastian Hellmann and Sebastian Nordhoff describe the Open Linguistics Working Group (OWLG), its goals, addressed problems, recent activities and on-going developments.

Christian Chiarcos describes the formalization of annotated linguistic corpora by means of OWL/DL with a focus on genericity and interoperability. Structural interoperability of linguistic corpora is addressed with POWLA, an OWL/DL formalization of a data model designed to represent any kind of linguistic annotation assigned to textual data; conceptual interoperability between annotations of different corpora can be established using the OLiA ontologies, an architecture of modular OWL/DL ontologies that formalize the linking of annotation schemes with community-maintained terminology repositories.

Sebastian Hellmann, Claus Stadler and Jens Lehmann describe the DBpedia, one of the major free data sets in the Web of Data, as an example of a lexical-semantic resource. In particular, the internationalization of the DBpedia is addressed – including the development of a German DBpedia. The authors further describes the NLP Interchange Format (NIF), that can be used, for example, to develop NLP pipelines that perform the task to assign words the corresponding DBpedia concept (entity linking). NIF represents the output of NLP tools in RDF, and thus, it is possible to integrate this data into an existing Linked Data infrastructure.

Sebastian Nordhoff presents a knowledge base that conveys information *about* linguistic resources, it thus exemplifies how metadata can be provided within the Linguistic Linked Open Data cloud: Sebastian Nordhoff describes how existing work on language classification can interface with bibliographical work based on standards like TEI and Dublin Core in the Glottolog/Langdoc project. His work affords links to the vast amounts of bibliographical data contained in the LOD cloud on the one hand, and language classification and language history on the other. Further, he illustrates the linking between LOD resources for the example of Glottlog/Langdoc and ASJP online, which measures the lexical distance between languages.

Using POWLA, the DBpedia, OLiA and Glottolog/Langdoc as examples, the final contribution by Christian Chiarcos, Sebastian Hellmann and Sebastian Nordhoff

describes how corpora, lexical-semantic resources, and other linguistic knowledge bases can be interlinked, and how additional information can be obtained by building a Linked Open Data (sub-)cloud of linguistic resources.

References

Baader F, Horrocks I, Sattler U (2005) Description logics as ontology languages for the Semantic Web. Mechanizing Mathematical Reasoning pp 228–248

Berners-Lee T (2006) Design issues: Linked data. http://www.w3.org/DesignIssues/LinkedData.html

Bird S, Liberman M (2001) A formal framework for linguistic annotation. Speech Communication 33(1-2):23–60

Bouda P, Cysouw M (this vol.) Treating dictionaries as a Linked-Data corpus. P. 15-23

Bray T, Paoli J, Sperberg-McQueen C, Maler E, Yergeau F (1997) Extensible markup language (xml). World Wide Web Journal 2(4):27–66

Burchardt A, Padó S, Spohr D, Frank A, Heid U (2008) Formalising Multi-layer Corpora in OWL/DL – Lexicon Modelling, Querying and Consistency Control. In: Proceedings of the 3rd International Joint Conference on NLP (IJCNLP 2008), Hyderabad

Cassidy S (2010) An RDF realisation of LAF in the DADA annotation server. In: Proceedings of the 5th Joint ISO-ACL/SIGSEM Workshop on Interoperable Semantic Annotation (ISA-5), Hong Kong

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

Declerck T, Lendvai P, Mörth K, Budin G, Váradi T (this vol.) Towards Linked Language Data for Digital Humanities. P. 109-116

Dipper S (2005) XML-based stand-off representation and exploitation of multi-level linguistic annotation. In: Proc. Berliner XML Tage 2005 (BXML 2005), Berlin, Germany, pp 39–50

Dostert L (1955) The georgetown-ibm experiment. In: Locke WN, Booth AD (eds) Machine Translation of Languages, John Wiley & Sons, New York, pp 124–135

Francis WN, Kucera H (1964) Brown Corpus manual. Manual of information to accompany A standard corpus of present-day edited American English, for use with digital computers. Tech. rep., Brown University, Providence, Rhode Island, revised edition 1979

Goldfarb CF, Rubinsky Y (eds) (1990) The SGML handbook. Oxford University Press, New York

Hellmann S (2010) The semantic gap of formalized meaning. In: The 7th Extended Semantic Web Conference (ESWC 2010), Heraklion, Greece

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

Ide N (1998) Corpus Encoding Standard: SGML guidelines for encoding linguistic corpora. In: Proceedings of the First International Language Resources and Evaluation Conference (LREC 1998), pp 463–70

Ide N, Pustejovsky J (2010) What does interoperability mean, anyway? Toward an operational definition of interoperability. In: Proc. Second International Conference on Global Interoperability for Language Resources (ICGL 2010), Hong Kong, China

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

Lassila O, Swick RR (1999) Resource Description Framework (RDF) model and syntax specification. Tech. rep., World Wide Web Consortium

McGuinness D, Van Harmelen F (2004) OWL Web Ontology Language overview. w3c recommendation. Tech. rep., World Wide Web Consortium

Miles A, Bechhofer S (2009) SKOS Simple Knowledge Organization System reference. W3C Recommendation. Tech. rep., World Wide Web Consortium

Morris W (ed) (1969) The American Heritage Dictionary of the English Language. Houghton Mifflin, New York

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

Schalley AC (this vol.) TYTO – A collaborative research tool for linked linguistic data. P. 139-149

Text Encoding Initiative (1990) TEI P1 guidelines for the encoding and interchange of machine readable texts. Tech. rep., Text Encoding Initiative, draft Version 1.1 1

Thompson S (1955-58) Motif-index of folk-literature: A classification of narrative elements in folktales, ballads, myths, fables, medieval romances, exempla, fabliaux, jest-books, and local legends. Indiana University Press, Bloomington

W3C OWL Working Group (2009) OWL 2 Web Ontology Language. document overview. W3C Recommendation. Tech. rep., World Wide Web Consortium

Wilcock G (2007) An OWL ontology for HPSG. In: Proc. 45th Annual Meeting of the Association for Computational Linguistics, Prague, Czech Republic, pp 169–172