Glottolog/Langdoc: Increasing the visibility of grey literature for low-density languages

Sebastian Nordhoff, Harald Hammarström

Max Planck Institute for Evolutionary Anthropology
Deutscher Platz 6, 04013 Leipzig, Germany
sebastian_nordhoff@eva.mpg.de, harald_hammarstroem@eva.mpg.de

Abstract

Language resources can be divided into structural resources treating phonology, morphosyntax, semantics etc. and resources treating the social, demographic, ethnic, political context. A third type are meta-resources, like bibliographies, which provide access to the resources of the first two kinds. This poster will present the Glottolog/Langdoc project, a comprehensive bibliography providing web access to 180k bibliographical records to (mainly) low visibility resources from low-density languages. The resources are annotated for macro-area, content language, and document type and are available in XHTML and RDF.

Keywords: Language Classification, Bibliography, Linked Data

macro-area	count	macro-area	count
Africa	53710	Pacific	13775
South America	21923	Eurasia	12263
North America	14880	Australia	8465

Table 1: Coverage by macro-area.

1. The myth of bad state of description

It is customary to deplore the unsatisfying documentary status of the world's language families (Comrie et al., 2005a; Krauss, 2007). It is true that for many of the world's languages, major Western publishers do not provide access to any form of documentation. This does not mean, however, that no documentation exists. (Hammarström and Nordhoff, 2011) showed that the percentage of languages with decent documentation is far greater than previously assumed, the problem is that the relevant works are often unpublished PhD theses, manuscripts, or books published by local agencies with no distribution channels in the Western world. This leads to the perception of a shortage of material, which is not necessarily the case.

2. Langdoc

2.1. Charting the descriptive status

The aim of the Glottolog/Langdoc project is to mobilize these resources, increase their visibility, and facilitate their discovery. For this purpose, we collected bibliographical records from >20 bibliographies with detailed coverage of particular areas. EBALL (Maho, 2010) for instance contains 50k references for Africa, (Fabre, 2005) contains 60k references for South America (of which not all are linguistic in nature). The challenge is to make these individual efforts available at a larger scale. At the time of writing, the aggregate of all bibliographies is 180k+ references focussing on low-density languages. The bibliographical ground work done by these dedicated individuals is exceptional.

2.2. Enriching

The source bibliographies differ in the kind and amount of detail they cover. In total, 83 different fields are used to store bibliographic information. While all input bibliographies list author, title, and year, the coverage of other interesting domains, such as language(s) discussed, document type (grammar, dictionary, text collection etc), or even the language the work is written in, varies. The three parameters just mentioned provide added value to a user in search of references for a given domain. We therefore enriched the initially sparsely populated fields with machine learning techniques.

The document type and the language described in the work are determined based on the title of the work (Hammarström, 2009; Hammarström, 2011). Typical titles are "A grammar of Lao" or "Wörterbuch der Nyakyusa-Sprache". The words found in titles fall into three categories: stopwords, words occurring in very many titles ('grammar', 'Wörterbuch'), and words occurring in very few titles ('Lao', 'Nyakyusa'). These sets can be automatically established based on informativeness. The very informative words normally refer to the language treated. (Hammarström, 2008) shows that about 70% accuracy in auto-annotation for language treated is achievable based on the title alone.

Auto-annotation for document type differs from auto-annotation for language in that there is only roughly a dozen document types, while there are thousands of languages with tens of thousands of names. However, parts of the source collection of documents have already been manually annotated as to document type, and Machine Learning techniques can be applied to generalize these annotations. (Hammarström, 2011) argues that a procedure similar to Decision Trees is the most appropriate technique due to some tricky dependencies between title words. Accuracy depends somewhat on the frequency and characteristics of individual labels, but F-scores in the range of 0.6-0.7 can be achieved.

Automatically inferring the language the work is written

reftype	count	reftype	count
grammar sketch	11349	text	816
ethnographic	9132	specific feature	813
grammar	8839	socling	596
dictionary	6352	dialectology	595
comparative	6261	bibliographical	526
wordlist	4266	minimal	463
overview	3992	new testament	137
phonology	1767		

Table 2: Coverage by reftype

in is the simplest of the annotation tasks and can be done at near 100%-accuracy by simply looking at frequent title words from each language in question (Hammarström, 2011).

This enriched annotation thus allows users to formulate very targeted queries such as 'Word list or dictionary from Nyakyusa written in German or English'.

2.3. Content search and browse

We have electronic copies of 7861 of the references in Langdoc. While we cannot make them available because of copyright issues, we do provide a fulltext search, which allows the user to further narrow down queries. Due to the rather low coverage of references available as full text (<5%), such queries fail to return a lot of legitimate references, but may nevertheless be useful.

To improve recall in keyword search we use query expansion with the k-nearest synonyms, where synonyms are automatically inferred using Random Indexing. Using Random Indexing both reduces the length of word-space vectors (and thus the computational cost of computing semantic similarity) and improves accuracy in predicting synonyms (Rosell et al., 2009).

For users who do not yet know which keywords they are interested in, we provide a browsing index, automatically generated from the fulltexts. The browsing index aims to list 'latent topics', i.e., topics that the references frequently deal with. In the present document collection, topics such as adjectives, pronouns, active/stative, ... may be expected, whereas very specific topics such as a language name or a specific morpheme in language should be avoided. An approach with TF-IDF weighting of document chunks has proven to achieve exactly this (Hammarström, 2012). We break down documents into chunks of section-length (0.5 to 5 pages in this text genre) yielding 110k document chunks. The top-n terms by TF-IDF are extracted from each chunk. This finds both latent topics and specific terms. To remove the highly specific terms, we may simply filter on frequency and keep those terms which occur (as the top-n TD-IDF terms) in many chunks. Table 3 shows such terms extracted (where $n = 10$) for the subset of fulltext references written in English.

We are currently experimenting with various forms of document classification to provide additional ways to browse Langdoc references.

term	#chunks	term	#chunks	term	#chunks
clause	106	sentence	63	stress	52
verb	94	object	63	relative	52
language	93	see	62	agreement	52
languages	89	subject	61	syllable	50
tone	81	linguistics	59	english	50
clauses	81	suffix	58	plural	49
university	80	nouns	58	phrase	49
class	79	rule	57	vowels	47
vowel	72	person	57	pronouns	47
noun	70	new	55	construction	46
verbs	68	chinese	55	base	46
case	66	malay	54	suffixes	45
press	65	high	53	roots	45
...	...				

Table 3: Terms occuring in the greatest number of document chunks.

3. Glottolog

The Langdoc repository is complemented by a repository of genealogical relations, Glottolog. Glottolog builds upon the classifications collected by the Multitree project[1] and contains 104 classifications with a combined total of 1 431 language families and 104 629 nodes. The main classification 'Glottolog 2012' contains 21 719 nodes in 431 families with a maximum depth of 19 levels. The 'Glottolog 2012' tree takes the 'Multitree Composite 2008' as a starting point, but has been thoroughly revised in accordance with specialist literature on individual families and subfamilies. This has lead to the rejection of many macro-families, such as "Altaic", and the many updated subgroupings have been implemented. The rejected families are retained as 'Spurious languoids' and provide links to the closest established languoids. As all other languoids, they have URIs.
The main classification furthermore responds to some additional constraints (unique names within classifications, names meaningful without context (no 'A' or 'Southern'), no one-member subfamilies, regularized treatment of chronolects like Latin).
Glottolog is tightly interlinked with Langdoc. There are 148 857 links between Langdoc references and the main classification 'Glottolog 2012'. When including all classifications, the number of links increases to 1 638 038. References are retrievable not only from the node they are attached to, but also from all higher nodes. This means that one can formulate queries like 'Give me all grammars of (((Central) East) Nuclear) Polynesian', next to maximally general queries like 'Give me all grammars of Austronesian' or maximally particular queries like 'Give me all grammars of Hawai'ian'. These genealogical queries can be combined with the bibliographical queries mentioned in Sect. 2..
This interlinking can be exploited for sampling purposes. A facility to draw a genetically and areally balanced sample of references of a certain type (grammar, dictionary, etc) is provided in the Glottolog/Langdoc interface. This procedure is fully automated and provides a pseudo-random sam-

[1] http://multitree.linguistlist.org

ple, which are of a higher sampling quality than the convenience samples often used in language typology (Nordhoff and Hammarström, 2011a).

Next to the main classfication, Glottolog contains additional classifications drawn from the Multitree project. Every node of these classifications has a unique ID, reflecting the insight that two researchers using the same name for a node do not always mean the same thing. The meaning of 'Altaic' for instance can be taken to include Korean and/or Japanese next to Turkic, Mongolic, and Tungusic. This means that the practice of the Multitree project to assign one and the same 4-letter code to all instances of Altaic (ALTC in this case) is not granular enough here. This is already evident from the list of alternate names Multitree gives for 'Altaic', among which we find 'Macro-Altaic' and 'Micro-Altaic', which clearly do not refer to the same entity. Glottolog assigns alphanumeric codes of the pattern `abcd1234` to all languoids, assuring maximal disambiguation possibilities.

4. Linked Data

All bibliographic records are treated as individual resources with their own URIs, as are all languages, dialects, and language families ('languoids'). These unique identifiers allow the integration of these resources into the semantic web (Nordhoff and Hammarström, 2011b) according to the principles of Linked Data (Berners-Lee, 2006; Heath and Bizer, 2011).

Glottolog makes use of concepts from the following ontologies: GOLD,[2] dcterms,[3] wgs84,[4] skos,[5] lexvo.[6] Additional concepts are provided in the glottolog ontology available at `http://glottolog.livingsources.org/ontologies/glottolog.owl`.

Glottolog/Langdoc is integrated into the emerging Linguistic Linked Open Data Cloud (`http://wiki.okfn.org/Wg/linguistics/llod`) (Chiarcos et al., 2012b; Nordhoff, 2012). (Chiarcos et al., 2012a) show in principle how a cross-domain query involving Glottolog/Langdoc and a set of annotated corpora can be formulated in SPARQL (Prudhommeaux and Seaborne, 2008). The query shown in Fig. 1 retrieves the labels of all syntactic categories associated with a languoid or any of its subnodes (assuming that corpora for the languages are available and that all corpora are annotated with glottolog languoid IDs). The amount of annotated corpora for low-density languages is of course lacking at the moment, but this example can still serve to illustrate the cross-domain interoperability of Linguistic Linked Data.

Next to granular accessibility, references can be downloaded in one bib-file, and all data can be downloaded as an rdf dump.

5. Modeling as RDF

Glottolog/Langdoc makes use of three basic concepts for the modeling of languoids and linguistic resources:

[2] `http://purl.org/linguistics/gold/`
[3] `http://purl.org/dc/terms/`
[4] `http://www.w3.org/2003/01/geo/wgs84_pos#`
[5] `http://www.w3.org/2004/02/skos/core#`
[6] `http://lexvo.org/ontology`

```
PREFIX glottolog: <http://glottolog.livingsources.org/ontologies/glottolog.owl\#>.
PREFIX dcterms:  <http://purl.org/dc/terms/>.
PREFIX powla:    <http://purl.org/powla/powla.owl\#>.
PREFIX olia:     <http://purl.org/olia/olia.owl\#>.
PREFIX rdfs:     <http://www.w3.org/2000/01/rdf-schema\#>.
CONSTRUCT { ?languoid <\#uses> ?syntacticCategory }
WHERE {
  ?languoid glottolog:sublanguoid ?sublanguoid
  ?node dcterms:language ?sublanguoid
  FILTER(regex(str(?languoid),"http://glottolog.livingsources.
                                org/resource/languoid/id/.*")).
  ?node a powla:Node.
  ?node a ?syntacticCategory
  FILTER(regex(str(?syntacticCategory),
              "http://purl.org/olia/olia.owl\#.*")).
  ?syntacticCategory rdfs:subClassOf olia:SyntacticCategory.
```

Figure 1: SPARQL query to retrieve all labels used for corpora of languages below a particular Glottolog node.

- a **lectodoc** is a document describing any kind of linguistic variety (any lect). Lectodoc is a subclass of `frbr:manifestation`.[7]
- a **doculect** is the linguistic system described in one lectodoc. Doculect is a subclass of `dcmi:linguisticSystem`.[8]
- a **languoid** is a set of doculects. It is a concept instantiated by doculects. Languoids can have sublanguoids and superlanguoids, just as any set can have subsets and supersets. For the time being, we are usings SKOS to model these relations.[9]

Glottolog/Langdoc provides URIs for every lectodoc, every doculect and every languoid.

The provision of URIs means that third parties can make use of Glottolog without the need to redo the whole project. Glottolog/Langdoc does for instance not believe in a node "Nostratic" and does not provide it in its main classification 'Glottolog 2012'. Given the availability of URIs for all top-level languoids (=language families), researchers who do not share this opinion could still publish RDF data stating the superset relationship between Nostratic and whatever language families they wish to include there.

```
myproject:Nostratic     glottolog:sublanguoid     glottolog:Indo1235
                                                 ,glottolog:Afro1235
                                                 ,glottolog:Kart1235
                                                 ,glottolog:Ural1235
                                                 ,….
```

Another use case is stating of identity of languoids from different authors. As discussed above, language families from different classifications cannot be assumed to refer to the same entity even if they happen to have the same name. This leads to a multiplication of URIs. There are close to 100k unique identifiers for languoids in Glottolog/Langdoc, far beyond the numbers usually used in linguistic typology. Many of these languoids are, however, very similar and can be considered very similar or identical in many (but not all) use cases. The definition of Basque is for instance uncontroversial, and every researcher agrees what should and should not be included in there. Glottolog/Langdoc does not equate these languoids, but other knowledge bases might want to assert `owl:sameAs`

[7] `http://purl.org/vocab/frbr/core`
[8] `http://purl.org/dc/terms/`
[9] `http://www.w3.org/2004/02/skos/core`

or `skos:closeMatch` between different authors' languoids with the name 'Basque', making use of the Glottolog/Langdoc URIs.
Given the document-based definition of languoids in Glottolog/Langdoc this can even be done automatically: every languoid which shares the same set of references with another languoid can be assumed to refer to the same entity in the real world. The existence of supplementary references in a languoid is more problematic, as this might be due to deeper coverage of the other languoid, or to a slightly broader coverage.

6. Use Cases

We can distinguish the following 6 use cases for Glottolog/Langdoc: 1) Query, 2) Browse, 3) Draw sample, 4) Compare, 5) Infer, and 6) Statistical Analysis.

6.1. Query

A user knowing what they are looking for can use a query mask to search for author, title, document type, languoid etc.

6.2. Browse

A user without a very specific query can browse Glottolog/Langdoc along the links provided inside the project and to related outside projects (Multitree,[10] LL-Map (Xie et al., 2009),[11] LinguistList,[12] Ethnologue (Lewis, 2009),[13] ODIN (Lewis, 2006),[14] WALS (Comrie et al., 2005b),[15] OLAC (Bird and Simons, 2001),[16] lexvo (de Melo and Weikum, 2008),[17] and Wikipedia.

6.3. Sample

A very specific use case is the fully automated, and therefore minimally biased, drawing of a sample. Such samples are for instance used in linguistic typology to test the worldwide distribution of linguistic features (Rijkhoff and Bakker, 1998; Bakker, 2011). Glottolog/Langdoc has the advantage that the sample of languages drawn can be constrained to languages having the required documentation while still being areally and genetically balanced. For instance, in order to do a comparison of phoneme inventories, one needs at least one of 'phonological description', 'sketch grammar' or 'grammar' for each language in the sample. Text collections will not do. Fully random sampling might return languages without the required documentation, which then have to be replaced. This is avoided in Glottolog/Langdoc. The sample will only return languages which respond to the selected criteria. If for some reason, the works relevant for a language cannot be procured, the 'replace lg' button allows to replace this language with its nearest genealogical neighbour responding to the selected criteria.

[10] `http://multitree.linguistlist.org`
[11] `http://www.llmap.org`
[12] `http://linguistlist.org/forms/langs/`
[13] `http://www.ethnologue.com`
[14] `http://www.csufresno.edu/odin/`
[15] `http://wals.info`
[16] `http://www.language-archives.org/`
[17] `http://www.lexvo.org`

lvl	AF	AUS	Eurasia	N.AM	PAC	S.AM	
4	588	121	527	249	358	247	2090
3	438	83	307	87	405	125	1445
2	126	25	156	32	168	53	560
<2	977	101	637	225	1099	175	3214
sum	2129	330	1627	593	2030	600	7309
avg	2.30	2.68	2.44	2.61	2.01	2.74	
%4	27.62	36.67	32.39	41.99	17.64	41.17	
%<2	41.66	12.42	33.68	29.17	52.41	6.50	

Table 4: Descriptive status of the areas of the world. The best available documentation determines the score: 4=grammar, 3=grammar sketch, 2=phonology or similar, 1=word list, 0=not even a word list published.

6.4. Compare

More advanced use cases are the graph-theoretic comparison of language classifications. Isomorphism of (sub)graphs of classifications by different authors can for instance be computed with Glottolog data. Another possibility are consensus trees.

6.5. Infer

A case of inference or automated reasoning was mentioned in the SPARQL query above (Fig. 1).

6.6. Statistical analysis

Finally, one can do statistical analysis of the density of coverage of a particular area or family (Hammarström and Nordhoff, 2011; Hammarström and Nordhoff, in press).
Statistic analysis of Glottolog/Langdoc data can give insights into the descriptive status of the languages of the world (Table 4). This table is to be read as follows: 588 African languages have a grammar as their most extensive piece of documentation, while 438 only have a grammar sketch (besides additional material irrelevant for this chart). This table does not count the quantity of documentation but the quality: 1 grammar will beat any number of wordlists. The other cells are to be interpreted in an analogous way. The table shows that about 2000 languages, or 28%, are at the highest level of documentation. About 3500 language, or roughly 50% have either a grammar or a grammar sketch. For 3200 languages, a wordlist is the best they can muster.

7. Theoretical implications

The interlinking of languoids and references and the good coverage of languoids allows us to change our definition of language. We have documents covering 7221 different languages in Glottolog/Langdoc, including some not included in ISO 639-3. Up to now, languages were defined by intension: language X is the language which is spoken there and there. We can now shift to an extensional definition: language X is what is described in the documents D, E, and F. This extensional definition has a number of advantages:

- **intersubjectivity**: researchers can easily agree on the identity of a document D. Agreeing on the identity of a language L is much more complicated.

- **computability**. The treatment of bibliographical references is well understood, and various tools for the handling of bibliographical data are available. Furthermore, bibliographical references are discrete,

class	subclass of	properties/remarks
Lectodoc	frbr:manifestation	hasdoculect
Doculect	dcmi:linguisticSystem	haslectodoc
Languoid	skos:concept	associatedDoculect, sublanguoid, superlanguoid
NonterminalLanguoid	Languoid	w/ sublanguoids
LanguageFamily	NonterminalLanguoid	
TerminalLanguoid	Languoid	w/o sublanguoids
LivingLanguage	TerminalLanguoid	
DeadLanguage	TerminalLanguoid	a cover term for languages not spoken today
ExtinctLineage	DeadLanguage	w/o offspring
Paleolect	DeadLanguage	w/ offspring
ProtoLanguage	Paleolect	reconstructed
ClassicalLanguage	Paleolect	directly attested

whereas languoids tend to have very fuzzy boundaries. Relying on discrete entities makes computation an easier task.

- **verifiability**: spurious claims about languages disappear. There are number of cases of languages with an ISO 639-3 code where it is absolutely unclear what these codes refer to (Nordhoff and Hammarström, 2011b). Taking documents as the basis of definition entails that one can always trace back where the claim to existence originated. Under the current 639-3 scheme, this is not always possible, as no sources are provided. Cases in point are the languages Cumeral [cum], Omejes [ome], Ponares [pod], and Tomedes [toe] supposedly spoken in Colombia, where it can not be ascertained whether they exist at all, but there existence cannot be disproved either as the basis for the claim to their existence is not disclosed by SIL, the ISO registrar.

The first and the second point above are intuitively clear. The third point deserves some more elaboration: Currently, researchers rely on ISO 639-3 codes to identify languages. The problem with ISO 639-3 is that the denotation of the codes is not always clear. For instance, the codes `ffi` (Foia Foia), `hhi` (Hoia Hoia), and `hhy` (Hoyahoya) refer to three Inland Gulf languages spoken in Papua New Guinea. But SIL, the ISO registrars, do not give any source for these three languages. As a result, it is impossible to ascertain whether Cridland's (1924) "Vocabulary of Mahigi" (Cridland, 1924), which clearly refers to an Inland Gulf language in the vicinity, would actually describe one of the three languages just mentioned, or whether it is an independent language. Under a document-centric, extensional approach, one could look up the documents the language is defined by and compare them with Cridland's treatise to evaluate whether this document can be assigned to any of the three languages. This approach scales nicely to higher levels of genealogical classification as well: The Inland Gulf languages can be described by the set union of `ffi`, `hhi`, `hhy` etc.

8. Conclusion

Glottolog/Langdoc provides URIs for languoids and documents which makes the domain of world wide linguistics fit for the semantic web and opens it up for a variety of use cases.

9. References

Dik Bakker. 2011. Language sampling. In Jae Jung Song, editor, *Handbook of Linguistic Typology*. OUP, Oxford.

Tim Berners-Lee. 2006. Linked data. http://www.w3.org/DesignIssues/LinkedData.html, 07.

Steven Bird and Gary Simons. 2001. The olac metadata set and controlled vocabularies. In *Proceedings of the ACL 2001 Workshop on Sharing Tools and Resources - Volume 15*, STAR '01, pages 7–18, Stroudsburg, PA, USA. Association for Computational Linguistics. Online version http://www.language-archives.org.

Christian Chiarcos, Sebastian Hellmann, and Sebastian Nordhoff. 2012a. Linking linguistic resources: Examples from the open linguistics working group. In Chiarcos et al. (Chiarcos et al., 2012c), pages 201–216. companion volume of the Workshop on Linked Data in Linguistics 2012 (LDL-2012), held in conjunction with the 34th Annual Meeting of the German Linguistic Society (DGfS), March 2012, Frankfurt/M., Germany.

Christian Chiarcos, Sebastian Hellmann, and Sebastian Nordhoff. 2012b. Towards a linguistic linked open data cloud: The open linguistics working group. *Traitement Automatique du Langage*.

Christian Chiarcos, Sebastian Nordhoff, and Sebastian Hellmann, editors. 2012c. *Linked Data in Linguistics. Representing Language Data and Metadata*. Springer. companion volume of the Workshop on Linked Data in Linguistics 2012 (LDL-2012), held in conjunction with the 34th Annual Meeting of the German Linguistic Society (DGfS), March 2012, Frankfurt/M., Germany.

Bernard Comrie, Matthew S. Dryer, David Gil, and Martin Haspelmath. 2005a. Introduction. In Bernard Comrie, Matthew S. Dryer, David Gil, and Martin Haspelmath, editors, *World Atlas of Language Structures*, pages 1–8. Oxford University Press.

Bernard Comrie, Matthew S. Dryer, David Gil, and Martin Haspelmath, editors. 2005b. *World Atlas of Language Structures*. Oxford University Press.

E. Cridland. 1924. Vocabulary of mahigi. *British New Guinea Annual Report*, 1923-1924:58–58.

Gerard de Melo and Gerhard Weikum. 2008. Language as a foundation of the Semantic Web. In Christian Bizer and Anupam Joshi, editors, *Proceedings of the Poster and Demonstration Session at the 7th International Semantic Web Conference (ISWC 2008)*, volume 401 of *CEUR WS*, Karlsruhe, Germany. CEUR.

Alain Fabre. 2005. Diccionario etnolingüístico y guía bibliográfica de los pueblos indigenas sudamericanos. Book in Progress at `http://butler.cc.tut.fi/~fabre/BookInternetVersio/Alkusivu.html` accessed May 2005.

Harald Hammarström and Sebastian Nordhoff. 2011. How many languages have so far been described? Paper presented at NWO Endangered Languages Programme Conference, Leiden, April 2011.

Harald Hammarström and Sebastian Nordhoff. in press. Achievements and challenges in the description of the languages of melanesia. In Marian Klamer and Nick Evans, editors, *Melanesian languages on the Edge of Asia*. Special Issue of Language Documentation & Con-

servation.

Harald Hammarström. 2008. Automatic annotation of bibliographical references with target language. In *Proceedings of MMIES-2: Workshop on Multi-source, Multilingual Information Extraction and Summarization*, pages 57–64. ACL.

Harald Hammarström. 2009. *Unsupervised Learning of Morphology and the Languages of the World*. Ph.D. thesis, Chalmers University of Technology and University of Gothenburg.

Harald Hammarström. 2011. Automatic annotation of bibliographical references for descriptive language materials. In Pamela Forner, Julio Gonzalo, Jaana Kekäläinen, Mounia Lalmas, and Maarten de Rijke, editors, *Proceedings of the CLEF 2011 Conference on Multilingual and Multimodal Information Access Evaluation*, volume 6941 of *LNCS*, pages 62–73. Berlin: Springer.

Harald Hammarström. 2012. Technologies for searching and browsing in descriptive grammars. Paper presented at the annual conference of the Graduate School of Language Technology (GSLT), Gothenburg, Sweden.

Tom Heath and Christian Bizer. 2011. *Linked Data - Evolving the Web into a Global Data Space*. Morgan & Claypool, San Rafael.

Michael E. Krauss. 2007. Mass language extinction and documentation: The race against time. In O. Miyaoka, O. Sakiyama, and M. Krauss, editors, *Vanishing Languages of the Pacific Rim*, pages 3–24. Oxford University Press.

W. D. Lewis. 2006. Odin: A model for adapting and enriching legacy infrastructure. In *Proceedings of the e-Humanities Workshop, held in cooperation with e-Science 2006: 2nd IEEE International Conference on e-Science and Grid Computing*. Amsterdam. online version available at http://www.csufresno.edu/odin/.

M. Paul Lewis, editor. 2009. *Ethnologue: Languages of the World*. SIL, Dallas, 16 edition.

Jouni F. Maho. 2010. User guide to EBALL. Document (version 19 May 2010) posted at `http://goto.glocalnet.net/eball/eballguide.pdf` accessed 3 March 2012.

Sebastian Nordhoff and Harald Hammarström. 2011a. Countering bibliographical bias with langdoc, a bibliographical database for lesser-known languages. Paper presented at ALT 9, Hong Kong, July 2011.

Sebastian Nordhoff and Harald Hammarström. 2011b. Glottolog/langdoc: Defining dialects, languages, and language families as collections of resources. In *Proceedings of ISWC 2011*.

Sebastian Nordhoff. 2012. Linked data for linguistic diversity research: Glottolog/langdoc and asjp online. In Chiarcos et al. (Chiarcos et al., 2012c), pages 191–200. companion volume of the Workshop on Linked Data in Linguistics 2012 (LDL-2012), held in conjunction with the 34th Annual Meeting of the German Linguistic Society (DGfS), March 2012, Frankfurt/M., Germany.

Eric Prudhommeaux and Andy Seaborne. 2008. Sparql query language for rdf. *W3C working draft*, 4(January).

J. Rijkhoff and D. Bakker. 1998. Language sampling. *Linguistic Typology*, 2-3:263–314.

Magnus Rosell, Martin Hassel, and Viggo Kann. 2009. Global evaluation of random indexing through swedish word clustering compared to the people's dictionary of synonyms. In *Proceedings of RANLP 2009*.

Yichun Xie, H. Aristar-Dry, A. Aristar, H. Lockwood, J. Thompson, D. Parker, and B. Cool. 2009. Language and location: Map annotation project - a GIS-based infrastructure for linguistics information management. In *Computer Science and Information Technology, 2009. IMCSIT '09. International Multiconference on*, pages 305 –311, oct. online version at http://www.llmap.org.