Harald HAMMARSTRÖM
Researcher, Max Planck Institute for Psycholinguistics (Netherlands)
(Stockholm, Sweden)

Glottolog: A Free, Online, Comprehensive Bibliography of the World's Languages

Glottolog (http://glottolog.org) is a *bibliography* of descriptive materials on the languages of the world (this part is also known as LangDoc) and a *classification* of the world's languages. Glottolog is browsable, searchable, downloadable, continually updated and free of charge.

The Glottolog bibliography was created in response to the lack of a sufficiently comprehensive and accessible bibliography on a world-level scale. Other large bibliographies exist already but have drawbacks on one or more of the desired aspects. For example, the SIL bibliography[74] accompanying the Ethnologue [Lewis et al. 2013] has a large number of bibliographical entries on lesser-known languages, but is restricted to work produced under the Summer Institute of Linguistics (SIL) umbrella. The Bibliographie Linguistique[75] is not restricted to a certain producer but systematically fails to include MA/PhD theses as well as items from minor countries, both of which make up a significant part of the total. Also, use of the Bibliographie Linguistique is not free of charge. Worldcat[76] has an enormous collection but also lacks large classes of items from major countries. Also, there is no systematic way of singling out linguistically relevant publications nor languages. Google Books[77] may have an even bigger coverage but, similarly, when it comes to lesser-known languages, there is no systematic way of singling out linguistically relevant publications nor languages.

The philosophy of comprehensiveness of the Glottolog bibliography is as follows:

A: Include the *most extensive pieces* (MED) of documentation for every language, and,

B: Beyond that, include "as much as possible".

[74] *Available at http://www.ethnologue.com/bibliography.asp.*

[75] *Available at http://bibliographies.brillonline.com/browse/linguistic-bibliography.*

[76] *Available at http://www.worldcat.org.*

[77] *Available at http://books.google.com.*

This implies that for a small language with only a wordlist to its documentation reference should be in Glottolog. For a bigger language with countless articles/books, a major dictionary/text/grammar collection should be included, but not necessarily every reference ever written about the language (but, of course, any amount of these are also welcome). In essence, only published or publicly accessible materials are considered, as opposed to ongoing unpublished work or manuscripts whose existence/access is difficult to confirm. Master's theses and PhD theses are included since they are in principle accessible from the approving institution.

Since there are over 7,000 languages in the world, the practice of actually obtaining the reference to the most extensive pieces of documentation for every language is highly non-trivial. In addition, language documentation and description is an extremely decentralized activity, carried out by missionaries, anthropologists, travellers, naturalists, amateurs, colonial officials, ethnographers and not least linguists over several hundred years. In fact, there has never in history been a systematic survey of descriptive materials on the languages of the world (although, at least, Adelung [1820] and Schmidt [1926] were in a position to produce one at times when the task was much smaller).

A legitimate question, then, is who knows all the obscure bibliographical references? The Glottolog answer is: experts on language families or areas know the bibliographical references for the corresponding families/areas. Experts write handbooks and overviews, such as outright bibliographies, comparative/descriptive overviews, and sociolinguistically oriented overviews. Following this logic, one may go through all handbooks and overviews and collect the references to obtain a comprehensive collection (for more details see [Hammarström and Nordhoff 2011]). The task of going through all handbooks/overviews is not necessarily a lesser amount of work because there are more handbooks/overviews than the number of languages (over 8,100, also listed in Glottolog and tagged as such). But it is more systematizable since countries, areas and families are easier to enumerate.

In addition to the strictly systematic collection, Glottolog also incorporates any existing bibliography available. A selection of the largest bibliographic databases granted by their respective compilers are listed in Table 1. A complete list with full descriptions of the source bibliographies and their provenance is available at http://glottolog.org/langdoc/langdocinformation. The total amount after removing duplicates is currently 193,407 references. This is not everything that has ever been written about any language, but the most extensive description for every language is included.

Table 1. Some existing bibliographical resources and their size, contents, annotation and the time the information was culled

	Number of references	Contents	Area	Coverage	Annotation		Date
EBALL	60,164	Everything	Africa	Full	100%	L & T	2009
HH	34,197	DD	World	85%?	100%	T	2014
Fabre	30,176	Everything	S. America	Full	100%	L	2009
SIL	18,464	Mainly DD & VP	World	70%?	100%	L & T	2009
MPIEVA	13,966	Everything	World	?	62-93%	L & T	2009
SILPNG	13,110	Mainly DD & VP	Papua	Full	100%	L & T	2004
ANLA	11,627	Mainly DD & MSS	Alaska	Full	100%	L	2012
OZBIB	10,377	Mainly DD	Australia	Full	100%	L	2010
WALS	5,633	Mainly DD	World	?	99%	L	2005

L = Language, T = Type, DD = Descriptive Data, VP = Vernacular Publications, MSS = Manuscripts

For enhanced sorting and filtering capabilities, references in Glottolog have some annotation. From the searcher's viewpoint, the more and the more detailed content-annotation the better, but from the annotators' viewpoint, more and more detailed annotation is more work, unless the annotation can be (semi-) automatized. In general, we only have access to the text of the bibliographical reference itself (author, title, year, etc.), not the actual document it refers to. Therefore, inferences depending on page counts or words that tend to occur in the title are possible, e.g., the name of the language(s) being treated often appears in the title (see below), but we cannot tell, e.g., whether there is a chapter/section on adjectives or whether numerals are included in a wordlist. As a compromise between search desiderata, annotation work and (semi-)automatizabillity, Glottolog references are annotated as to language and description type. As to language, references are tagged with ISO-639-3 language code(s)[78] which allows lookup for location, speaker numbers, etc. Ideas on annotation also for other levels (above the language level, i.e., a

[78] *See http://www.sil.org/iso639-3/default.asp.*

(sub)family or below the language level, i.e., a (sub)dialect) are currently being implemented (cf. [Cysouw and Good 2013]). As to description type, Glottolog references are annotated according to the hierarchy in Table 2. Roughly half of the references in Glottolog are manually annotated, often by translation of an annotation scheme used by a source bibliography, and the other half is automatically annotated based on words in the title of the reference. Essentially, a reference titled "A grammar of Tauya" can be inferred to be of the description type grammar and the language Tauya (see [Hammarström 2008, 2011] for details). Since the automatic annotation has a much higher error rate than manual annotation, it is recorded for every reference which annotation is automatic and which is manual, and this can be used for filtering. However, quality assessments requiring specialized knowledge fall outside the current scope of Glottolog.

Table 2. The typology of description types used in Glottolog

Type	Explanation
grammar	a description of most elements of the grammar ~ 150 pages and beyond
grammar sketch	a less extensive description of many elements of the grammar ~ 50 pages
dictionary	~ 75 pages and beyond
text	text material
specific feature	description of some elements of grammar (i.e., noun class system, verb morphology, etc.)
wordlist	~ a couple of hundred words
minimal	A small number of morphemes
overview	Document with meta-information about the language (i.e., where spoken, non-intelligibility to other languages, etc.)

The Glottolog website has entry points for simple searches on references, simple searches on languages and (sub)families and complex searches for reference and (sub)family at the same time. In the latter way one can, e.g., search for all grammars for African languages of the Semitic subfamily produced in 1984.

An example of a typical Glottolog view is shown in Figure 1 focussing on the language Sakha (also known as Yakut). The tree showing the classification of Sakha is shown at the top left and the geographical location on the map on the right. The tree is navigable and the references listed below provide the justification for why Sakha is (sub)classified the way shown. The classification

employs an even standard of evidence required across all families/languages/areas of the world[79]. Since the evidence for most families can be debated to some degree, pointers to the arguments supporting each node are given in brief with references to literature along with comments if needed. The list at the bottom contains the bibliographical references tied to Sakha. The list can be filtered, sorted and downloaded in various formats.

Figure 1. *An example of a typical Glottolog view*

The entire Glottolog database, both the classification and the bibliography, can be downloaded[80] along with older versions. Versions starting from Glottolog 2.3 are long time archived with a DOI[81]. The data is also available as Linked Open (see [Forkel 2014]).

References

1. Adelung, F. (1820). *Uebersicht aller bekannten Sprachen und ihrer Dialekte*. St.Petersburg: Nic. Gretsch.
2. Cysouw, M. & Good, J. (2013). Languoid, Doculect, Glossonym: Formalizing the notion "language". *Language Documentation and Conservation* 7. 331–359.

[79] *For more detailed information, see: glottolog.org/glottolog/glottologinformation.*

[80] *From the downloads page http://glottolog.org/meta/downloads.*

[81] *See http://dx.doi.org/10.5281/zenodo.*

3. Forkel, R. (2014). The Cross-Linguistic Linked Data project. In: C. Chiarcos, J. P. McCrae, P. Osenova & C. Vertan (eds.), *3rd Workshop on Linked Data in Linguistics: Multilingual Knowledge Resources and Natural Language Processing*, 60–66. Reykjavik, Iceland: European Language Resources Association (ELRA).
4. Hammarström, H. (2008). Automatic annotation of bibliographical references with target language. In: *Proceedings of MMIES-2: Workshop on Multi-source, Multilingual Information Extraction and Summarization*, 57–64. ACL.
5. Hammarström, H. (2011). Automatic Annotation of Bibliographical References for Descriptive Language Materials. In: P. Forner, J. Gonzalo, J. Kekäläinen, M. Lalmas & M. de Rijke (eds.), *Proceedings of the CLEF 2011 Conference on Multilingual and Multimodal Information Access Evaluation* (LNCS 6941), 62–73. Berlin: Springer.
6. Hammarström, H. & Nordhoff, S. (2011). LangDoc: Bibliographic Infrastructure for Linguistic Typology. *Oslo Studies in Language* 3(2). 31–43.
7. Lewis, P. M., Simons, G. F. & Fennig, C. D. (2013). *Ethnologue: Languages of the World*. 17th edn. Dallas: SIL International.
8. Schmidt, W. (1926). *Die Sprachfamilien und Sprachenkreise der Erde* (Kulturgeschichtliche Bibliothek. Reihe 1, Ethnologische Bibliothek 5). Heidelberg: Carl Winter's Universitätsbuchhandlung.