LANGDOC: BIBLIOGRAPHIC INFRASTRUCTURE FOR LINGUISTIC TYPOLOGY

HARALD HAMMARSTRÖM AND SEBASTIAN NORDHOFF
Nijmegen and Leipzig

ABSTRACT

The present paper describes the ongoing project **LangDoc** to make a bibliography website for linguistic typology, with a near-complete database of references to documents that contain descriptive data on the languages of the world. This is intended to provide typologists with a more precise and comprehensive way to search for information on languages, and for the specific kind information that they are interested in. The annotation scheme devised is a trade-off between annotation effort and search desiderata. The end goal is a website with browse, search, update, new items subscription and download facilities, which can hopefully be enriched by spontaneous collaborative efforts.

[1] INTRODUCTION

Language Typology is the subfield of linguistics concerned with *the systematic study of the unity and variation of the languages of the world*. Like many disciplines, there are various infrastructural needs which are not yet in place. A central such need is as follows. Typically, the material for study for a typologist is a document with descriptive information on a language. With some 7 000 languages in the world (Lewis 2009), the number of relevant such documents grows far beyond the capacity of individual typologists. At present, single individuals have to manage micro-collections of references for their own use, which means not only gathering and re-typing them but also performing very time-consuming searches. The present paper describes a project **LangDoc** aimed at eradicating this enormous duplication of work, by providing a free and (if not complete) extensive collection of bibliographical references[1] available for download, search, subscription etc via a website.

In essence, the goal of LangDoc is as follows:

---

[1] Many of the actual documents that the references point to are difficult to access, tucked away only here and there in libraries across the world. Arguably, there is a similar superfluous duplication of work involved in accessing them. However, the present paper does not address this matter, which appears to be vastly more complicated than collecting only the references.

- Delineate a class of bibliographical references, namely those to descriptive materials

- Annotate them with focus (what language, family, etc.) and with type (wordlist, phonology, grammar etc.) such that

    - basic search criteria are met

    - the identity- and type-annotation has good automatization prospects

- Provide an updateable website interface

We will first define the scope of the proposed collection of references, and discuss some existing databases. Next, we will address issues of annotation and search desiderata. Finally, we will touch issues of update management, community contribution and crediting.

[2]   THE SCOPE OF THE BIBLIOGRAPHICAL DATABASE

[2.1]   *Desired Scope*

At present, a bibliography of all relevant research articles, e.g., 'all articles ever written in linguistics' or even 'all articles relevant for typology', however useful, seems much too large to be feasible. However, a bibliography of *descriptive* materials of the languages of the world is a fairly well-delineable class. For short, a bibliographic reference to a publication with descriptive/documentational data and/or metadata (number of speakers, location etc) will be called a **BDP.** The class of BDPs, as opposed to a mix of general linguistics articles, is of salient usefulness for a typologist. Furthermore, albeit with some work, it appears to be within scope to achieve a (near-)complete such database the following sense:

A) For every language, include the most extensive piece(s) of documentation, *and*

B) Beyond that, include "as much as possible"

This policy implies that

- for a small language with only a wordlist to its documentation, that BDP should be included

- for a bigger language with countless articles/books, a major dictionary/ grammar/text collection should be included, but not necessarily every single BDP ever written about the language (but, of course, any amount of these are also welcome)

[2.2] *Collecting References*

Language documentation and description is, and has been, an extremely decentralized activity. For well over two centuries, there has been intensive collection of data on the languages of the world by missionaries, anthropologists, travellers, naturalists, amateurs, colonial officials, and not least linguists. For natural reasons, all these people, including the linguists, hail from all parts of the world and call from maximally disconnected research environments. As a result, finding and tracking references to descriptive materials is not a straightforward task.

Traditionally, bibliographies would be curated by individual researchers, often experts on some area or language family, who happened to take on the matter after decades of collection, and then published in book form. These, when available and recent, are excellent guides, but do not cover the entire world (unless accumulated – see below), which is usually the frame of interest of the typologist. There are also a few bibliographies which have world-wide scope, but which are imperfect to the needs of the typologist in one or the other way. For example, the Ethnologue website[2] by SIL International lists references, but almost all of them are to works by SIL affiliated authors – a significant but small subset of the entire author space – and systematically excludes languages that went extinct before 1950, even if they are well-documented. The Linguistic Bibliography Online website[3] systematically fails to include MA/PhD theses and items from minor countries, and requires a subscription fee. The Worldcat catalogue[4] also fails to include many MA/PhD theses and other items for minor countries, and has no way of singling out linguistically relevant publications. Though some entries in Worldcat have annotation, overall, this is so unsystematic that it is of little use for finding BDPs on, e.g., a small Papuan language. Google, Google Scholar, and Google Books are, of course, resources with enormous coverage, but for browsing or zooming in on a specific language or area, it is difficult to come up with high-precision searches.

Now, given how decentralized language description is, one may doubt why it should even be possible to build a bibliographical database that meets high standards of completeness and precision. Who knows of all the obscure BDPs? We submit that experts of countries/language families/areas do tend to know the BDPs, obscure and non-obscure, of their respective field of interest. These experts write overviews and handbooks on a regular basis. For example, one type of overview with BDPs is a traditional printed book bibliography, such as:

Newman, Paul. (1996) *Hausa and the Chadic Language Family: A Bibliography.*

---

[2]   http://www.ethnologue.com accessed 1 Jan 2010. The printed edition in book form does not have all the references that the website has.

[3]   http://www.blonline.nl/public/ accessed 1 Jan 2010. Printed editions in book form appear annually.

[4]   http://www.worldcat.org accessed 1 Jan 2010.

Köln: Köppe [African Linguistic Bibliographies 6].

Another type of overview is a descriptive overview, i.e., an overview of what languages there are and a little about their nature in a certain area, such as:

Laycock, Donald C. (1968) *Languages of the Lumi Subdistrict.* Oceanic Linguistics VII(1):36-66.

Further, perhaps the most common kind of overview with bibliographical references to the languages covered is a historical-comparative work, such as:

Adam, Lucien. (1893) *Matériaux pour servir à l'établissement d'une grammaire comparée des dialectes de la famille Caribe.* Paris: J. Maisonneuve [Bibliothèque Linguistique Américaine XVII].

In addition, there are sociolinguistically oriented overviews, such as:

Shearer, Walter and Sun Hongkai 2002 *Speakers of the Non-Han Languages and Dialects of China*, Lewiston, NY: Edwin Mellen Press [Chinese Studies 20]

and so on. Thus, going through all such overviews and handbooks collecting the references, is a systematic procedure for attaining a satisfactory world-wide bibliographical database. However, this only holds if there exist (recent) experts covering the whole world and that all their handbooks and overviews can be enumerated, since they, too, are of the same decentralized nature as the descriptive works on the languages themselves. The difference is that there are much fewer experts, areas, families and countries than there are languages, so the matter is more manageable. Nevertheless, the absolute number of overviews exceeds 5 000, according to our own collections so far.

[2.3]   *Some Existing Resources*

Related to the above questions of how to collect and what to collect, significant headways have already been made in the actual work of doing the collection. Table 1 lists some existing resources of special interest to the present project.

All the resources of Table 1 are updated regularly, wherefore we report the time the information was collected. The Electronic Bibliography of African Languages and Linguistics (EBALL)[5] by Jouni Filip Maho, the Diccionario Etnologüístico y Guía Bibliográfica de los Pueblos Indígenas Sudamericanos (here abbreviated DEPIS)[6] by Alain Fabre, World Grammar Bibliography (WGB) by Harald Hammarström are bibliographies collected by single dedicated individuals following

---

[5]   See http://goto.glocalnet.net/maho/eball.html, accessed 1 Jan 2010. WEB-BALL by Guillaume Ségérer is an online query interface that is based on an independently updated earlier version of EBALL (with ca 50% of the entries of the 2009 version) available at http://sumale.vjf.cnrs.fr/Biblio/index.html accessed 1 Jan 2010.

[6]   See http://butler.cc.tut.fi/~fabre/BookInternetVersio/Alkusivu.html accessed 1 Jan 2010.

	# Refs	Contents	Area	Coverage	Annotation		Date
EBALL	60 164	Everything	Africa	Full	100%	L & T	Sep 2009
DEPIS	30 176	Everything	S America	Full	100%	L	Sep 2009
WGB	15 103	DD	World	85%?	100%	T	Dec 2009
MPIEVA	13 966	Everything	World	?	62-93%	L & T	Sep 2009
WALS	5 633	Mainly DD	World	?	99%	L	Aug 2005
SIL	18 464	Mainly DD & VP	World	70%?	100%	L & T	Sep 2009
SILPNG	13 110	Mainly DD & VP	Papua	Full	100%	L & T	Sep 2004

TABLE 1: Some existing bibliographical resources and their size, contents, annotation and the time the information was culled. Abbrevations are L = Language, T = Type, DD = Descriptive Data, VP = Vernacular Publications.

more or less the methodology outlined above; to go through all overviews. While EBALL and DEPIS strive to include everything, not just BDPs, on the respective languages, including all references to work done on relatively well-studied languages (such as Aymara or Hausa) and including non-descriptive work where the language in question is brought up (for example, in a discussion of the merits of a linguistic theory), WGB only strives to include the best descriptive work(s) on every language. This is the reason WGB has worldwide scope but is much smaller than the respective area-specialist bibliographies. MPIEVA is the online queryable library catalog of the Max Planck Institute for Evolutionary Anthropology[7]. In contrast to many other libraries, there is a dedication to collect descriptive data on the languages of the world, and most of the entries are annotated with `ISO 639-3` codes, which makes it relatively simple to extract the part of the catalogue which refers to descriptive works. The WALS is a landmark multi-person typological project whose bibliography is `ISO 639-3` annotated and available on the web[8]. The SIL Bibliography is the bibliography[9] of the missionary/linguist organization SIL International whose members have worked on a significant part of the world's lesser described languages. SILPNG (Akerson & Moeckel 1992; Linden 2003; Feldpausch 2005a,b) is a paper bibliography of the Papua New Guinea branch of SIL, where a significant part of the world's lesser described languages are found. SIL is a decentralized organization, and not all SILPNG references are included in the SIL Bibliography.

Access and license matters to the above collections are not yet clear, but it is likely that all of them can be used for benevolent purposes.

---

[7] http://www.eva.mpg.de/english/library.htm accessed 1 Jan 2010.

[8] http://wals.info/refdb/search accessed 1 Jan 2010

[9] http://www.ethnologue.com/bibliography.asp accessed 1 Jan 2010.

[3.1]   *Baseline Functionality*

Essentially, the typologist is looking for a BDP either from the language-side or from the document-contents-side (or a combination). Searching from the language side is typically to get whatever references are associated with a particular language, or associated with the language(s) that have some property such as 'belonging to family X' or 'endangered'. From the document-contents-side, the typologist may be looking for kinds of content of the document, such as 'contains wordlist', 'contains a section on adjectives' or 'contains interlinear glossed text'.

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 and 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.

Based on experience, the authors propose the following annotation scheme as a compromize between search desiderata, annotation work and (semi-)automatizabillity.

Identity: The language(s) the BDP treats. As a baseline, we suggest ISO 639-3[10] codes should be used as the identity registry. ISO 639-3 codes are preferable as a baseline since linguists are used to them and they have good automatization properties. Furthermore, there already exists a database from which location, speaker number, genealogical classification etc. can be retrieved from ISO 639-3 codes.

Other identity schemes, notably the doculect-languoid scheme (Cysouw & Good 2007; Good & Hendryx-Parker 2006) are more dynamic, and will in the end supersede the special status of the level of a maximal set of mutually intelligible varieties, which is the backbone of the ISO 639-3 division (Lewis 2009, 7-18). For this reason, we also foresee a complementary, open-ended, identity annotation scheme which allows arbitrary (groups of) varieties on the sub-language level.

Type: The type/content of the document the BDP refers to. As a midway between our impression of typologists search desiderata, already existing annotation (e.g., from library catalogues) and (semi-)automatizability, we propose the following relatively uncontroversial hierarchy:

---

[10]   See http://www.sil.org/iso639-3/default.asp accessed 1 Jan 2010

- (full-length) descriptive grammar
- grammar sketch
- dictionary
- description of some element of grammar (i.e., noun class system, verb morphology etc)
- phonological description
- text (collection)
- wordlist
- document with meta-information about the language (i.e., where spoken, (non-)intelligibility to other languages etc.)

We wish to stress the importance of partial automatizability of BDP annotation, which is some kind of guarantee that the endeavor will actually lead to a finished product and that updates are not very expensive.

As an example of how partial automatization of BDPs may work, we walk through an experiment described in Hammarström (2008) on how ISO 639-3 language identity codes may be extracted from the title line of a BDP.

More formally, the problem may be cast as follows:

Given: A database of the world's languages (consisting minimally of <unique-id, language-name>-pairs)

Input: A bibliographical reference to a work with descriptive language data (= a BDP) of (at least one of) the language in the database

Desired output: The identification of which language(s) is described in the bibliographical reference

Unfortunately, the problem is not simply a clean database lookup! For example, a BDP might look as follows:

> Dammann, Ernst. (1957) *Studien zum Kwangali: Grammatik, Texte, Glossar.* Hamburg: Cram, de Gruyter & Co [Abhandlungen aus dem Gebiet der Auslandskunde / Reihe B, Völkerkunde, Kulturgeschichte und Sprachen 35].

This reference happens to be written in German. In general, the metalanguage could be any language (ca. 30 actually occur). The reference happens to describe a Namibian-Angolan language called Kwangali, ISO 639-3 *kwn* and the task is to automatically infer this using a database of the world's languages and/or databases of other annotated bibliographical entries, but without humanly tuned thresholds. In the ISO 639-3 database, each language has a three letter id, a canonical name and a set of variant and/or dialect names, for example

$Words(e_t)$	$LN(Words(e_t))$	$Words(e_t)$	$LN(Words(e_t))$
étude	{}	cameroun	{}
du	$\{dux\}$	du	$\{dux\}$
samba	$\{ndi, ccg, smx\}$	nord	{}
leko	$\{ndi, lse, lec\}$	famille	{}
parler	{}	adamawa	{}
d'allani	{}		

TABLE 2: For an example entry e_t, we show how many ISO 639-3 identities are associated with each word in the title of the entry.

> Canonical name: Kwangali
> ISO 639-3: kwn
> Alternative names: {Kwangali, Shisambyu, Cuangar, Sambio, Kwangari, Kwangare, Sambyu, Sikwangali, Sambiu, Rukwangali}.

The languages and language name database consists of 7 299 languages, 42 768 language name tokens, 39 419 unique name strings. It is not yet well-understood how "complete" this language name database is, but as a rough indication we manually checked 100 randomly chosen bibliographical entries, whose titles contained a total of 104 language names. 43 of these names (41.3%) existed in the database as written, and 66 (63.5%) existed in the database, if one allows for spelling variation.

The size of the language name database is both a blessing and a burden. It may first seem as simple as looking up every word in the title of a BDP and pick the language whose name matches at least one word. Unfortunately, such a procedure only gets around 20% accuracy. To see why, consider the following example BDP:

> Fabre, Anne Gwenaïélle. (2002) *Étude du Samba Leko, parler d'Allani (Cameroun du Nord, Famille Adamawa).* PhD Thesis, Université de Paris III – Sorbonne Nouvelle.

The ISO 639-3 codes whose language name matches at least one word in the title is shown in Table 2. It so happens that such a common strings of letters as *du* happens to be a language name! The correct classification is this case is only $\{ndi\}$.

Clearly, we cannot guess blindly which word(s) in the title indicate the target language. But we can exploit some domain specific properties:

- A title of a publication in language description typically contains

    (i) One or few words with very precise information on the target language(s), namely the name of the language(s)

	foe	pole	huli	papua	guinea	comparativen	
$WC(w)$	1	2	3	57	106	110	
Rel.Inc.	1.0	2.0	1.5	19.0	1.86	1.04	
	new	study	languages	and	a	the	of
...	145	176	418	1001	1101	1169	1482
...	1.32	1.21	2.38	2.39	1.10	1.06	1.27

TABLE 3: The values of $WC(w)$ for w taken from an example entry (mid row). The bottom row shows the *relative increase* of the sequence of values in the mid-row, i.e., each value divided by the previous value (with the first set to 1.0).

(ii) A number of words which recur throughout many titles, such as 'a', 'grammar', etc.

- Most of the languages of the world are poorly described, there are only a few, if any, publications with original descriptive data.

Thus a more clever way is to divide the words in the title into two groups, informative and non-informative, and only use the informative ones for lookup. How can we measure the informativeness of a word w? Let $WC(w)$ = the number of distinct codes associated with w in the training data (set of already annotated BDPs) or Ethnologue database. Then for each word w, we get a value of informativeness. The question remains, at which point (above which value?) of informativeness do we get a near-unique language name rather than a relatively ubiquitous non-informative word? Luckily, we are assuming that there are only those two kinds of words, and that at least one near-unique language will appear. Thus, if we cluster the values into two clusters, the two categories are likely to emerge nicely. The simplest kind of clustering of scalar values into two clusters is to sort the values and put the border where the relative increase is the highest. The following example illustrates the method:

W. M. Rule 1977 *A Comparative Study of the Foe, Huli and Pole Languages of Papua New Guinea*, University of Sydney, Australia [Oceania Linguistic Monographs 20]

Table 3 shows the title words and their associated number of codes (sorted in ascending order).

The highest relative increase is 19.0 between Huli and Papua. Thus, Foe, Pole and Huli are deemed near-unique and the rest non-informative. In this example, the three near-unique identifiers are correctly singled out.

The above method achieves about 70% accuracy, which can be slightly improved by allowing for spelling variants and disambiguation schemes (for details see Hammarström 2008).

So far we have not experimented with type-annotation, but impressionistically a similar level of accuracy seems achievable.

[3.2]   *Extended Functionality*

Slightly more challenging than browsing for document properties is the browsing of language family trees. Depending on the scope of the research question, speech varieties smaller or bigger than the traditional 'language' are of interest. For instance, dialectologists will find it useful to narrow down their searches to the dialects of Croatian spoken in Italy instead of stopping at the language level of 'Croatian' ISO 639-3 hrv and be provided with information about Standard Croatian and other irrelevant dialects. On the other hand, comparatists will find it useful to have a node of all Scandinavian Northern Germanic languages together instead of having to collect the references for each language separately (ISO 639-3 swe, ISO 639-3 nor, ISO 639-3 dan, etc). This is even more relevant for less well-known language families and large-scale typology, where queries like "Give me a reference to every full description of a Nilotic language" are perfectly normal. It is therefore interesting to go beyond the flat list provided by ISO 639-3 and add information about genetic nodes above and below the level of language as defined by the ISO-codes.

Existing genetic linguistic classifications can be exploited for this purpose. The multitree-project[11] contains a number of different linguistic classifications of the languages of the world in XML-format. Among these are so-called 'composite trees', which combine classifications of one family by different authors, diverging in scope and detail, into a much larger tree. These composite trees contain information about dialects as well as overarching large family classifications on a continental scale. A language typologist can select a node on the tree which corresponds to the scope of his or her study (dialect, language, language family, or any level in between). This node can then be used in database queries, together with the BDP properties mentioned above. A query on a node will return all documents which are attached to the node itself or any of its daughter nodes.

A major problem is that the assignment of BDPs to arbitrary nodes is more difficult to automatize than the assignment of BDPs to the standardized set of 7589 ISO-language names. For the time being we aim at attaching all BDPs to nodes which have an ISO-code as a start. Chosen users will be granted the right to reassign BDPs to other nodes interactively in a browser interface. Most typically, this will mean assigning a particular BDP to a subvariety below the node with the ISO-code, e.g.

---

[11]   http://linguistlist.org/multitree/ accessed 1 Jan 2010.

Sammartino, Antonio. (2004) *Grammatica della lingua Croato-Molisana.* Zagreb: Fondazione "Agostina Piccoli".

would be reassigned from the node `<node name="Croatian" iso639-3="hrv">` to `<node name="Molise Croatian" iso639-3="">`. This graphical user interface will also allow users to add new BDPs and to assign them to the relevant nodes, assuring that the project will go with the times.

[4] ORGANIZATION AND MANAGEMENT

As already declared, the goal of LangDoc is a website with a comprehensive and annotated BDP bibliography with functionality such as browsing, searching, updating, new items subscription and downloading. BDPs have a well-defined structure and there are no interesting technical aspects of providing a web-interface to them.

At present, a functioning such website is not far away. However, it is useful to also consider how to best keep it updated, and how to make it a functioning collaborative resource. To encourage the submitting of additions/corrections by the public, and to give credit where credit is due, the information on who submitted the entry should be saved and displayed. Another option is to allow major resources to be "published" under the website's umbrella, with a clear identity surrounding it. The advantage of putting it under the umbrella would be that it is integrated in tools and search scopes of the overarching website.

[5] CONCLUSION

The present paper describes LangDoc, a project to make a bibliography website for linguistic typology, with a near-complete database of references to documents that contain descriptive data on the languages of the world. This provides typologists with a more precise and comprehensive way to search for information on languages, and for the specific kind information that they are interested in. The annotation scheme devised is a trade-off between annotation effort and search desiderata. In addition to saving time, such a database also has other uses. For example, there are so far unanswered questions about exactly how many and which languages of the world have been described, which have not, and which have partial descriptions. Another use has to do with the growing uneasiness of typologists towards the notion of language as a maximal set of mutually intelligible varieties. The typologist may also be interested in sub-language-level varieties and contrast between them, and may therefore want to build a catalogue of varieties (rather than languages). Such a catalogue of varieties is naturally based on the target documents of BDPs, and defining a variety reduces to saying which BDPs fall within it.

REFERENCES

Akerson, P. & B. E. R. Moeckel. 1992. Bibliography of the Summer Institute of Linguistics Papua New Guinea Branch 1956-1990. In *Summer Institute of Linguistics.* Ukarumpa, Eastern Highlands Province, Papua New Guinea.

Cysouw, M. & J. Good. 2007. Towards a comprehensive languoid catalogue. Presentation at the TOWARDS A COMPREHENSIVE LANGUOID CATALOGUE workshop at the MPI for Evolutionary Anthropology, Leipzig, 28 June 2007. Available at http://email.eva.mpg.de/~haspelmt/cat.html.

Feldpausch, B. 2005a. Bibliography of the SIL Papua New Guinea Branch 2001-2003 including materials produced by the Bible Translation Association of Papua New Guinea in cooperation with SIL Ukarumpa, Eastern Highlands Province, Papua New Guinea: Summer Institute of Linguistics .

Feldpausch, B. 2005b. Bibliography of the SIL Papua New Guinea Branch 2004 including materials produced by the Bible Translation Association of Papua New Guinea in cooperation with SIL Ukarumpa, Eastern Highlands Province, Papua New Guinea: Summer Institute of Linguistics .

Good, J. & C. Hendryx-Parker. 2006. Modeling contested categorization in linguistic databases. In *Proceedings of the EMELD 2006 Workshop on Digital Language Documentation: Tools and Standards: The State of the Art.* Lansing, Michigan. http://www.linguistlist.org/emeld/workshop/2006/papers/GoodHendryxParker-Modelling.pdf.

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

Lewis, P. M. (ed.). 2009. *Ethnologue: Languages of the world.* Dallas: SIL International, 16th edn. Online Version: http://www.ethnologue.com.

Linden, L. 2003. Bibliography of the Summer Institute of Linguistics Papua New Guinea Branch 1991-2000. Ukarumpa, Eastern Highlands Province, Papua New Guinea: Summer Institute of Linguistics.

AUTHOR CONTACT INFORMATION

Harald Hammarström
Centre for Language Studies
Radboud Universiteit
Postbus 9103
NL-6500 HD Nijmegen
The Netherlands

Max Planck Institute for Evolutionary Anthropology
Deutscher Platz 6
D-04103 Leipzig
Germany
h.hammarstrom@let.ru.nl

Sebastian Nordhoff
Max Planck Institute for Evolutionary Anthropology
Deutscher Platz 6
D-04103 Leipzig
Germany
sebastian_nordhoff@eva.mpg.de