Linked Data for Linguistic Diversity Research: Glottolog/Langdoc and ASJP Online

Sebastian Nordhoff

Abstract Most of the linguistic resources available to day are about the world's major languages. This paper discusses two projects which have world-wide coverage as their aim. Glottolog/Langdoc is an attempt to attain near-complete bibliographical coverage for the world's lesser-known languages (i.e. 95% of the world's linguistic diversity), which then provides solid empirical ground for extensional definitions of languages and language classification. Automated Similarity Judgment Program (ASJP) online provides standardized lexical distance data for 5800 languages from Brown et al (2008) as Linked Data. These two projects are the first attempt at a Typological Linked Data Cloud, to which PHOIBLE by Moran (this vol.) can easily be added in the future.

1 Introduction

The original motivation underlying the development of standards such as RDF has been to describe resources, e.g. books in a library. The Glottolog/Langdoc project exemplifies a similar application scenario for the linguistic domain, i.e. the collection and formalization of **information about languages and language resources** within the Linked Open Data cloud. By doing so, Glottolog/Langdoc covers the band-width of languages in the world as far as possible, i.e. with a certain emphasis – albeit not a strict focus – on less-resourced languages.

Section 2 gives an overview of the bibliographical part of the project (Langdoc), Sect. 3 introduces the notion of **languoid**, a data structure for the modeling of genealogical relationships between language families, languages and dialects (Glottolog), Sect. 4 summarizes the resource types provided for the Linguistic Linked

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

Open Data cloud, and Sect. 5 illustrates a concrete application of Glottolog/Langdoc in the context of the related ASJP project.

2 Multilingual References on a World-wide Scale: Langdoc

Linguistic resources can be classified into resources *for* a language (dictionaries, thesauri, spellcheckers etc) and resources *about* a language (descriptive grammars, description of the history of a language, theoretical linguistic analysis of the phonology/morphology/syntax of a language). Linked Open Data has made good progress in the former area (WordNet etc), but has not really started yet in the latter. The Langcoc project aims at remedying this by providing near-complete bibliographical information about the world's lesser known languages as Linked Data.

2.1 Lesser-Known Languages

Current estimates suggest that there are about 7000 different languages spoken on Earth. As far as the amount of resources available for a language is concerned, there is a clear split. One the one hand, we have high resource languages. Those are national languages (e.g. Swedish), or other languages with a long written tradition (Catalan). These languages provide commercial viability for linguistic resources. On the other hand, we have languages with a very short written tradition, or even no written tradition at all. For those languages, there is no commercial interest in providing linguistic resources. The resources treating those languages are academic or missionary. The Langdoc project focuses on the latter group. I will call this group of languages 'lesser known languages'. Other names for the same group are 'low-density languages', and 'low resource languages'. Our current estimate is that there are only about 200 'better known languages' and the remainder, i.e. 6800 languages, have to be considered lacking in terms of resources.

2.2 Resource Collection and Resource Collections

A number of dedicated individuals have consecrated a lot of their time to collecting references about lesser-known languages. Alain Fabre (Fabre, 2005) has collected 26 634 references, treating 615 languages of South America; Jouni Maho (Maho, 2001) has collected 59 788 references treating 1 994 languages of Africa. *SIL international* have collected 6 246 references for Papua New Guinea (410 languages), in addition to the 18 190 references they provide on a world-wide scale. The Alaska

Native Center[1] provides 13 876 references for the languages of Alaska. The coverage of these bibliographies can be considered near-complete. For the other areas of the world, there are no comparable bibliographies. We have to rely on the aggregation of a number of large world-wide bibliographies and a number of smaller areal bibliographies[2] and hope that they will complete each other. It is likely that the more obscure references are currently not included in our resource collection for those areas of the world.

Langdoc lists 166 459 resources providing information about the world's linguistic diversity.[3] The resources are tagged for resource type (grammar, word list, text collection etc), macroarea (geographic region), and language. Table 1 gives an overview of the resources covered so far, classified by macroarea and document type.

Table 1 Language resources in Langdoc according to geographic region and document type

area	refs	document type	refs	document type	refs
Africa	74 787	comparative treatise	13 827	phonology	1 942
South America	32 897	grammar sketch	13 810	bibliography	1 464
Eurasia	16 879	ethnographic treatise	13 504	specific feature	1 362
Pacific	15 424	grammar	10 209	text	1 039
Australia	7 557	overview	8 273	sociolinguistics	943
North America	3 815	dictionary	7 408	dialectology	797
Middle America	1 897	wordlist	5 552	new testament	143

Our plan is to significantly expand the coverage of Langdoc in the years to come. The ingestion of future resources is guided by the following two principles, where the first has a higher priority than the second:

---

[1] http://www.uaf.edu/anla/

[2] ASJP Automated Similarity Judgment Program bibliography http://lingweb.eva.mpg.de/asjp/index.php/ASJP; Alain Fabre's "Diccionario etnolingüístico y guía bibliográfica de los pueblos indígenas sudamericanos" http://www.tut.fi/ fabre/BookIntervetVersio; The bibliography of the Papua New Guinea branch of SIL http://www.sil.org/pacific/png/; Randy LaPolla's Tibeto-Burman bibliography http://victoria.linguistlist.org/ lapolla/bib/index.htm; The bibliography of the South Asian Linguistics Archive http://www.sealang.net/library/; Frank Seifart's bibliography www.eva.mpg.de/lingua/staff/seifart.html; The World Atlas of Language Structures www.wals.info; Harald Hammarström's bibliography http://haraldhammarstrom.ruhosting.nl/; The catalogue of the Max Planck Institute for Evolutionary Anthropology in Leipzig, www.eva.mpg.de/library; The SIL bibliography www.ethnologue.com/bibliography.asp; The web-version of EBALL, by Jouni Maho and Guillaume Ségerer http://sumale.vjf.cnrs.fr/Biblio/; Jouni Maho's bibliography of Africa; http://goto.glocalnet.net/maho/eball.html; Tom Güldemann's bibliography of Africa http://www2.hu-berlin.de/asaf/Afrika/Mitarbeiter/Gueldemann.html; Chintang-Puma Documentation Project http://www.uni-leipzig.de/∼ff/cpdp/

[3] Note that the we only provide the reference, but no copy of the work itself. We link to WorldCat, GoogleBooks and Open Library to help users retrieve a copy.

1. For every language, provide a reference of the most extensive piece of documentation.
2. Beyond that, provide as many references as possible

2.3 Storage and Retrieval

All references are stored in a relational database. A web frontend provides access to the data. The references are retrievable via standard bibliographical fields such as author, year, title, etc. Additionally, Langdoc allows for for genealogical searches in a step-free manner. This is accomplished by using a set-theoretic approach: English is a subset of Germanic, and a subset of Indo-European. This means that a reference associated with English is associated with Germanic (and Indo-European) at the same time.

A researcher interested in languages of the Pacific Ocean could search at any level of the deeply nested tree of Austronesian languages (Fig. 1). Queries like 'Give me any dictionary of an Oceanic language' or 'Give me any grammar of a Polynesian language' become possible. The genealogical data just mentions lead us to the counterpart of Langdoc: Glottolog.

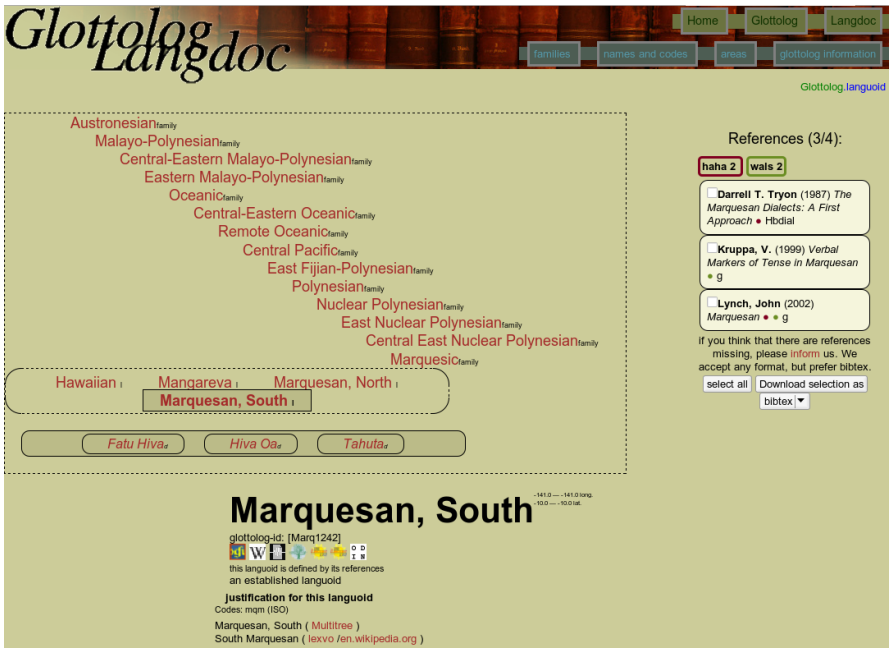

Fig. 1 The page for the languoid 'Marquesan' with the genealogy on the left and references on the right.

3 A Resource-Based Definition of Languages: Glottolog

The traditional approach to language classification has an intensional approach: languages have an essence. The problem is that this essence is not accessible to computers, so that models relying on the essence of a particular language are difficult to implement. To circumvent this issue, we use a novel approach in that we use an *extensional* definition of languages: a language is defined by the set of documents which describe it (Nordhoff and Hammarström, 2011). This has two advantages: we can exploit the available understanding of the ways how to model documents and associated metadata, and languages without documentation disappear from the model space. While it would of course be desirable to have information about any and all languages ever spoken by humans, it is a fact that we do not have all this information, and the scientific method dictates that we stick to what is observable. In this sense, the disappearance of languages without empirical attestation from our model is actually an advantage. This does not mean that those languages would be less interesting; it simply means that they are not yet part of the observed universe of Western academia.

Another advantage is that relations between languoids can be modeled in a set theoretic fashion. Let $\mathfrak{S}$ be the set of all documents treating Swedish, $\mathfrak{D}$ be the set of all documents treating Danish and $\mathfrak{N}$ be the set of all documents treating Norwegian. The union of these sets is then the set of all documents treating a North Germanic language. This subset relation can be iterated up to the root node 'Indo-European'.

This set-theoretic approach allows for a step-free modeling of language classifications. It furthermore does away with the need of singling out a particular level 'language' (as compared to dialect or variety), which short-circuits eternal discussions as to whether variety X is a language or a dialect. All sets have persistent IDs, which can be used to uniquely refer to a set, and no IDs are privileged.

The provision of URIs for documents and sets means that conflicting opinions can be modeled: Some researchers for instance assume that Baltic and Slavic are direct children of Indo-European while others assume an intervening node Balto-Slavic. The unique URIs allow researchers disagreeing about this aspect of the classification to state that the lower parts of the tree would still be identical. The document-centric approach furthermore allows the inference that Baltic, Slavic, and Indo-European still have the same meaning in both classifications, based on the extensional definition based on documents. The only differences is that in the second classification, there will be an additional languoid (more on this term below) 'Balto-Slavic' with associated documents, which is entirely missing from the first classification.

This leads to the modeling employed by Glottolog. As stated above, we employ a set-theoretic approach. Every languoid is seen as a set. Subset and superset relations can model genealogical relationships. In this particular case, Glottolog employs `skos:narrower` and `skos:broader` to model the relation between a larger languoid like North Germanic and a smaller languoid like Danish. Treating languages as concepts allows to make use of general insights gained in other areas where taxonomies of concepts are used. At the same time, this means that Glot-

tolog/Langdoc languoids are not comparable to languages in the sense of GOLD (Farrar and Langendoen, 2003a) or Ethnologue (Lewis, 2009b). The former are concepts (as defined in SKOS) while the latter are linguistic systems (as defined in Dublin Core).

4 Glottolog/Langdoc and Linked Data

Glottolog/Langdoc provides two types of resources as Linked Open Data: languoids and bibliographical records.

Languoid is a cover term for dialect, language, and language family (Good and Hendryx-Parker, 2006). Every languoid has its own URI and is annotated for ancestors, siblings, children, names, codes, geographic location and references. Links are provided to Multitree,[4] LL-Map (Xie et al, 2009) , LinguistList,[5] Ethnologue (Lewis, 2009a), ODIN (Lewis, 2006), WALS (Dryer, 2005), OLAC (Bird and Simons, 2001), lexvo (de Melo and Weikum, 2008), and Wikipedia. Languoids are modeled using SKOS and RDFS and linked to ontologies like GOLD (Farrar and Langendoen, 2003b), lexvo, and geo,[6].

Bibliographical records of **Language resources** are available in XHTML and RDF. Resources make use of Dublin Core (Weibel et al, 1998) and are annotated for the languoids they are applied to. Additionally, resources are linked to WorldCat,[7] GoogleBooks,[8] and Open Library.[9]

The database currently covers 166 459 resources and 94 049 languoids. Tables 1 and 2 provide an overview of its content according to different criteria.

The data collected within the Glottolog/Langdoc project are different from other linguistic data and cannot be linked according to the standard principles (e.g. *lemon*, see McCrae this volume). There are three nodes in the Linked Data cloud Glottolog/Langdoc can be hooked onto: ISBN numbers can link Glottolog/Langdoc to various resources, ISO 639-3 codes can link Glottolog/Langdoc to lexvo, and the provided geocodes can link to geographical data repositories like WGS84.[10]

[4] `http://multitree.linguistlist.org`
[5] `http://linguistlist.org`
[6] `http://www.w3.org/2003/01/geo/wgs84_pos`
[7] `http://www.worldcat.org`
[8] `http://books.google.com`
[9] `http://openlibrary.org`
[10] `http://www.w3.org/2003/01/geo/wgs84_pos`

language	refs	language	refs	language	refs
Swahili	1 916	Igbo	550	16 languages	300-399
Hausa	1 609	Sotho, Southern	539	31 languages	200-299
Nama	1 288	Arabic, Algerian	526	159 languages	100-199
Zulu	1 060	Oromo, Borana-Arsi-Guji	516	389 languages	50-99
Arabic, South Levantine	1 033	Turkish	511	647 languages	25-49
Yoruba	925	Tarifit	505	611 languages	15-24
Kabyle	897	Nyanja	504	533 languages	10-14
Thai	745	Arabic, Tunisian	498	1033 languages	5-9
Pulaar	743	Tachelhit	490	351 languages	4
Xhosa	739	Wolof	487	436 languages	3
Akan	729	Tibetan	483	612 languages	2
Éwé	713	Sotho, Northern	467	1045 languages	1
Tswana	703	Aymara, Central	462		
Mapudungun	610	Aymara, Southern	454		
Shona	597	Vietnamese	439		
Somali	591	Paraguayan Guaraní	436		
Amharic	554	Singa	405		

Table 2 Documentation status of the languages in Langdoc (excluding resource-heavy languages like English or German). The relative overrepresentation of African and South American languages is due to the extent of bibliographical coverage found Maho's (Maho, 2001) and Fabre's (Fabre, 2005) work.

5 Lexical Distances as Linked Data Resources: ASJP Online

Having described the Glottolog/Langdoc resources, we will now turn to an example application that integrates these resources with lexical-semantic resources provided by another project, ASJP (Holman et al, 2011).

Genetic relatedness between languages can be established by comparing basic vocabulary from the languages under discussion in order to see whether cognate sets and corresponding sound changes can be found. This has been a very laborious task for a very long time. The Automated Similarity Judgment Program (ASJP) automatizes this task by providing standardized word lists for 5 395 languages according to a particular abstract phonetic representation (Brown et al, 2008). Using a distance metric like Levenshtein, the relative lexical distance between two languages can be computed. The results can be compiled in a distance matrix and be represented as a tree. Figure 2 shows a screenshot illustrating this for Slavic languages.

The ASJP website currently also allows on-the-fly clustering and dendrogram generation of custom sets of languages. A researcher interested in the classification of Basque could for instance compute a tree of a candidate family and Basque to see how Basque fits into this family as far as its basic vocabulary is concerned.

The resources of this project are currently being made available as Linked Data in RDF[11] including language names, ISO 639-3 codes, WALS codes, number of speakers, date of extinction (if applicable), longitude, latitude, 40-items word list, and lexical distance between any two languages. Language names, codes, and geo-

[11] `http://cldbs.eva.mpg.de/asjp`

ASJP
languages in family SLAVIC

id	name	iso	wals	speakers	longitude	latitude
1603	BELARUSIAN	bel	blr	8618630	28.00	54.00
1604	BOSNIAN	bos	bos	2203800	18.00	43.00
1605	BULGARIAN	bul	bul	9097220	25.00	42.50
1606	CROATIAN	hrv	scr	5546590	15.98	45.82
1607	CZECH	ces	cze	9490840	15.00	50.00
1608	KASHUBIAN	csb	ksu	3000	18.00	54.00
1609	LOWER_SORBIAN	dsb	srl	7240	14.33	51.75
1610	LOWER_SORBIAN_2	dsb	srl	7240	14.33	51.75
1611	MACEDONIAN	mkd	mcd	2113170	21.75	41.67
1612	NINILCHIK_RUSSIAN	rus	rus	143553950	-151.54	59.99
1613	OLD_CHURCH_SLAVONIC	chu		-2	22.95	40.63
1614	POLISH	pol	pol	39990670	20.00	52.00
1615	RUSSIAN	rus	rus	143553950	38.00	56.00
1616	SERBOCROATIAN	srp	scr	7020550	19.00	44.00
1617	SLOVAK	slk	svk	5019950	20.00	49.00
1618	SLOVENIAN	slv	slo	1909050	15.00	46.00
1619	UKRAINIAN	ukr	ukr	37029730	33.00	49.00
1620	UPPER_SORBIAN	hsb	sou	18240	14.50	51.83

show family in Glottolog show family in Ethnologue

Fig. 2 The ASJP page for Slavic languages. A list of the languages on the left is complemented by an unrooted tree showing the lexical proximity of the Slavic languages. We can distinguish East Slavic languages on top, South Slavic Languages to the left, and West Slavic languages to the right and at the bottom.

graphical coordinates allow for the integration of ASJP data into the Linked Open Data cloud. For instance, Glottolog makes use of an ASJP as a provider for dendrograms for arbitrary nodes.

ASJP links to the following other language projects (Glottolog/Langodoc, Ethnologue, OLAC, Multitree, LL-MAP, lexvo) via `dcmi:relation`. Further research will be necessary to provide a predicate more information than `dcmi:relation`. ASJP has incoming links from Glottolog/Langdoc.

ASJP provides lexical information in a wider sense. This information can in principle be linked to other lexical semantic resources like WordNet. This is however complicated by the particular phonetic representation ASJP uses as a storage format. This representation makes use of a limited set of characters. For instance, the German word *Knochen* 'bone' is represented as <knoX3n>. More serious is the reduction of phonological oppositions. As an example, roundedness is never recorded in ASJP lexemes. This means that French *deux* 'two' is represented in ASJP as <de>, the same representation *dé* 'dice' would receive. The conversion from Standard French phonology to ASJP representation is straightforward: /ø/ and /e/ both translate to <e>. The inverse direction is more problematic: <e> corresponds to both /ø/ and /e/, and it is not clear how an ASJP string like <de> should be translated. This phonological underspecification means that automated linking to other language-particular lexical semantic resources is severely hampered.

The third and last type of data are the lexical distances computed for pairs of languages. These distances are made available as floating point numbers. This means that third party projects can use these data for other purposes, e.g. refinement of the clustering algorithm or similar.

References

Bird S, Simons G (2001) The olac metadata set and controlled vocabularies. In: Proceedings of the ACL 2001 Workshop on Sharing Tools and Resources - Volume 15, Association for Computational Linguistics, Stroudsburg, PA, USA, STAR '01, pp 7–18, DOI http://dx.doi.org/10.3115/1118062.1118065, URL http://dx.doi.org/10.3115/1118062.1118065, online version http://www.language-archives.org

Brown CH, Holman EW, Wichmann S, Velupillai V (2008) Automated classification of the world's languages: A description of the method and preliminary results. STUF 61(4):286–308

Dryer MS (2005) Genealogical language list. In: Comrie B, Dryer MS, Gil D, Haspelmath M (eds) World Atlas of Language Structures, Oxford University Press, pp 584–644

Fabre A (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.

Farrar S, Langendoen D (2003a) A linguistic ontology for the semantic web. Glot International 7(3):97–100

Farrar S, Langendoen D (2003b) Markup and the GOLD ontology. In: EMELD Workshop on Digitizing and Annotating Text and Field Recordings, Michigan State University

Good J, Hendryx-Parker C (2006) Modeling contested categorization in linguistic databases. In: Proceedings of the EMELD Workshop on Digital Language Documentation, East Lansing, Michigan

Holman EW, Brown CH, Wichmann S, Müller A, Velupillai V, Hammarström H, Sauppe S, Jung H, Bakker D, Brown P, Belyaev O, Urban M, Mailhammer R, List JM, Egorov D (2011) Automated dating of the world's language families based on lexical similarity. Current Anthropology 52:841–875

Lewis M (ed) (2009a) Ethnologue: Languages of the World, Sixteenth edition. SIL International, Dallas, online version available at http://www.ethnologue.com/. Accessed on 2011-11-27.

Lewis MP (ed) (2009b) Ethnologue: Languages of the World, 16th edn. SIL, Dallas

Lewis WD (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, URL http://faculty.washington.edu/wlewis2/papers/ODIN-eH06.pdf, online version available at http://www.csufresno.edu/odin/

Maho J (2001) African Languages Country by Country: A Reference Guide, Göteborg Africana Informal Series, vol 1, 5th edn. Department of Oriental and African Languages, Göteborg University

de Melo G, Weikum G (2008) Language as a foundation of the Semantic Web. In: Bizer C, Joshi A (eds) Proceedings of the Poster and Demonstration Session at

the 7th International Semantic Web Conference (ISWC 2008), CEUR, Karlsruhe, Germany, CEUR WS, vol 401

Moran S (this vol.) Using Linked Data to create a typological knowledge base. P. 129-138

Nordhoff S, Hammarström H (2011) Glottolog/Langdoc: Defining dialects, languages, and language families collections of resources. In: Proceedings of ISWC 2011, URL http://iswc2011.semanticweb.org/fileadmin/iswc/Papers/Workshops/LISC/nordhoff.pdf

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

Xie Y, Aristar-Dry H, Aristar A, Lockwood H, Thompson J, Parker D, Cool B (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, pp 305 –311, DOI 10.1109/IMCSIT.2009.5352710, URL http://ieeexplore.ieee.org/stamp/stamp.jsp?tp=&arnumber=5352710, online version at http://www.llmap.org