Glottolog/Langdoc: Defining dialects, languages, and language families as collections of resources

Sebastian Nordhoff & Harald Hammarström

Max Planck Institute for Evolutionary Anthropology

Abstract. This paper describes the Glottolog/Langdoc project, an attempt to provide near-total bibliographical coverage of descriptive resources to the world's languages. Every reference is treated as a resource, as is every "languoid"[1]. References are linked to the languoids which they describe, and languoids are linked to the references described by them. Family relations between languoids are modeled in SKOS, as are relations across different classifications of the same languages. This setup allows the representation of languoids as collections of references, rendering the question of the definition of entities like 'Scots', 'West-Germanic' or 'Indo-European' more empirical.

1 Dialects, languages, and language families

The question of what is a dialect and what is a language is a very old one, and up to now, there are no agreed upon criteria how to resolve it. While it is a hotly debated topic among the general public, there is general consensus among linguists that this question is of relatively minor interest. The classical quotation summarizing the problems of defining a language was popularized by Max Weinreich: "A language is a dialect with an army and a navy". This highlights the socio-political dimensions of declaring something a 'dialect' or a 'language'. To give an illustration: Before the break-up of the former Yugoslavia, Serbo-Croatian was considered a single language, whereas now Bosnian, Croatian, and Serbian are considered three distinct languages despite their grammars not having undergone any change. The reason for this change in status is clearly political and not linguistic.

On the level of language family, the disputes are less important, but disagreement still exists. Some linguists argue for instance that Quechua is a language family comprising 2, 6, or 46 languages, while others argue that Quechua is one language with a certain number of dialects. Political considerations also play a role here: a pan-Quechuan identity advocated by the Academia Mayor de la Lengua Quechua is easier to vindicate if they share a common language rather than if they share a common language family.

The difficulties of defining a language have a direct bearing on Semantic Web applications. ISO 639-3[1] is a standard for identifying languages in electronic resources. This works fine as long as this standard governs user preferences,

[1] `http://www.sil.org/iso639-3/`

locales and the like. But when it comes to automatic reasoning, ISO 639-3 is actually insufficient due to the lack of a clear definition of the languages it assigns codes to. For instance, German as spoken in Northern Germany does not have a distinction between /eː/ and /æː/, while in the South, this distinction can be found. When pulling together resources about the language with the code 'deu' (German), counts of the number of vowels will be conflicting. This is not due to different analyses, but rather to different beasts being analyzed. URIs relying on 'deu', 'eng', 'fra' etc thus do not identify exactly one resource, but rather point to a random member of a set of resources. Western languages are often standardized, the obvious example being French, but such standardization is only found for a small percentage of the 7000+ languages with an ISO 639-3 code. For languages without a written standard, the problems alluded to above in the German example become much more serious. Standardization furthermore does normally not refer to phonological issues like the distinction between /eː/ and /æː/.

ISO 639-3 is a good choice when dealing with the major languages of the world, but it is not granular enough for the Semantic Web when world wide linguistic diversity is the research topic, as for instance in the field of linguistic typology. We address this problem by using a resource-based definition of linguistic varieties ('languoids'), based on collections of resources treating them. Unique identifiers are provided for languoids at every level, allowing a step-less differentiation where required. Differences in categorization are also taken care of by tying URIs to the namespace of their creators, so that German as defined by Smith might exclude Swiss German, but German as defined by Miller might include it.

2 Definitions

We use the following technical terms for our purpose

- a **lectodoc** is a document containing information about a linguistic variety (a lect). This can be a grammar, a dictionary, a word list, or sociodemographic information about the linguistic group ("The ABC live on the XYZ river as hunter-gatherers and speak DEF").
- a **doculect** is a documented lect, i.e. a linguistic variety described in exactly one lectodoc. This means that on this level of analysis, every grammar of German represents a (slightly) different doculect, by virtue of being a different document. Very similar doculects will typicall be grouped into a languoid.
- a **languoid** [1] is a collection containing doculects or languoids. This recursive definition means that every languoid is reducible to doculects, which are anchored in lectodocs, identifiable resources.
- **Glottolog** is a project providing information about and URIs for 60k languoids
- **Langdoc** is a project providing bibliographical data and URIs for 200k lectodocs

– **Glottolog/Langdoc**[2] is the connection of Glottolog and Langdoc URIs. It can be found at http://glottolog.livingsources.org

These terms and their roles in the Glottolog/Langdoc project will now be explicated in more detail.

3 Lectodocs and Doculects

To reuse a foundational metaphor of linguistics: lectodocs and doculects are two sides of the same coin. A lectodoc is a document, e.g. a book, which describes a doculect, e.g. Catalan. It is impossible to dissociate one and the other. It is true that speakers of Catalan do not need a document to know that they are speaking Catalan and to make use of their language, but here we are dealing with linguistic research on a world-wide scale, which crucially depends on documents storing information about languages. The doculect is thus the thing described, the content, while the lectodoc is the container of the description, the document. As a shorthand test: a doculect is something you can speak while a lectodoc is something you can print. While this distinction is not crucial for many linguists, in the context of Linked Data, it is very important in order to not pollute the Semantic Web with illicit inferences.

In the Glottolog/Langdoc project, lectodocs are annotated for author, year, title etc using vocabulary from Dublin Core[2] and BIBO.[3] Lectodocs are further annotated in the Glottolog/Langdoc namespace for 'document-type' (grammar, dictionary, wordlist, etc). Lectodoc-URIs make use of numerical identifiers since existing identifiers like ISBN do not cover our whole document space. For instance, M.A. theses from Brazil do normally not have ISBN numbers.

4 Languoids

Languoids replace the traditional concepts of dialects, languages, and language families in the Glottolog/Langdoc project. Languoids are mathematically sets, which can contain other languoids, or doculects. Languoids may not be the empty set.

As an illustration, take the languoid l9267 'Kamu' (a language of Australia). This is defined as the set of the doculects {d278503}, a singleton set in this case. The leading 'd' of the only member stands for doculect and the digits refer to the ID of the corresponding lectodoc, in this case Harvey, M. (1990) *A sketch grammar of Gamu*.

The sister language Madngele is defined by the doculects contained in the lectodocs of the set {d282482}, again a singleton. d282482 can be dereferenced as Zandvoort, F. B. (1999) *A grammar of Matngele*.

Higher languoids are represented by sets of languoids. As an example, the mother of the two languoids just mentioned is l9266, Eastern Daly. It is defined as

[2] `http://dublincore.org/`
[3] `http://bibliontology.com/`

the union of the two daughter languoids: {l9267,l9268}. In this case, the leading 'l' stands for 'languoid' and the digits refer to the ID of the languoid.[4]

Languoids are annotated for their children (sublanguoids) with `skos:narrower`,[5] parents (superlanguoids) with `skos:broader` and immediately contained lectodocs with `rdfs:isDefinedBy`. Languoids are furthermore annotated for common names and codes. One crucial aspect of Glottolog/Langdoc is that languoids have creators, i.e. linguists who have proposed a given languoid. This languoid then resides in their namespace. Tying languoid names to creators allows us to disentangle common confusions with regard to higher groupings. The string 'Dravidian' is a common moniker for a language family of South Asia, but the internal structure of the tree can vary, as the following two partial trees show

(1) Ethnologue 2005:Dravidian = {Central Dravidian, Northern Dravidian, South-Central Dravidian, Southern Dravidian, Unclassified Dravidian}

(2) Zvelebil 1970:Dravidian = {Central Dravidian, North Dravidian, South Dravidian, Telugu, Toda-Badaga, Tuḷu}

Central Dravidian, *North(ern) Dravidian* and *South(ern) Dravidian* occur in both superlanguoids. But the two superlanguoids have different numbers of sublanguoids, 5 and 6 respectively, making them not isomorphic on a graph theoretic level. The labels *Unclassified*, *South-Central*, *Telugu*, *Toda-Badaga*, and *Tuḷu* are also not shared, a further indication that the semantics of *Dravidian* as conceived by Zvelebil and *Dravidian* as conceived by the Ethnologue cannot be equated.

What can be done is to state that the two representations are similar, given that both creators had the shared goal of representing those languages spoken in South Asia which are related to Tamil, Telugu, Kannada etc. The URIs provided by Glottolog/Langdoc allow to formulate relations between two languoids, e.g. as `skos:closeMatch`, which has the definition: "used to link two concepts that are sufficiently similar that they can be used interchangeably in some information retrieval applications"[6] with the emphasis on *some*. An example would be

(3) `glottolog:l109398 skos:closeMatch glottolog:l2712 .`

This kind of predication would not be possible with ISO 639-3 because ISO 639-3 does not provide a code for Dravidian, and ISO 639-3 would never distinguish between two different representations of the same language family. It is true that ISO 639-5[7] provides some labels for language families, but these are by no means exhaustive, and do not allow us to distinguish between different creators, which is crucial as shown for the Dravidian example above.

[4] It is also possible to have sets with both languoids and doculects, but this case is rare. It arises for instance when a work treats a language family as a whole. This will not be developed in detail here.

[5] `http://www.w3.org/2004/02/skos/`

[6] `http://www.w3.org/TR/2009/CR-skos-reference-20090317/#mapping`

[7] `http://www.loc.gov/standards/iso639-5/`

It has been stated above that the two trees/graphs of Dravidian are not isomorphic and can therefore not be arguments of `owl:sameAs`. However, it is clear that the two types of Dravidian are closer to each other than either is to a tree of Austronesian for instance. This can be modeled by reducing the tree representations to flat lists of lectodocs attached below the nodes to be compared. If the two sets of lectodocs are identical, `owl:EquivalentClass` is warranted since the concepts have the same extension. If one is a proper subset of the other, more computation is necessary, taking into account other trees. This will not be discussed here, but will often result in applying `skos:closeMatch`.

5 Advantages

The Glottolog/Langdoc approach has a number of advantages:

- URIs are given for every languoid, where other systems like ISO 639-3/5 restrict themselves to languages and selected families
- The availability of unique identifiers at every level allows to circumvent the issue of singling out a more prestigious level of 'language', short-circuiting some perpetual discussions.
- Namespaces for authors assure that different languoids are not lumped together if the name is the same but the semantics differ.
- Every languoid has to be grounded in a resource; languoids without resources (which are commonly found in Ethnologue, e.g. Loarki[8]), are not permitted. This means that intersubjectivity can be attained: two researchers can make sure that they are talking about the same thing, which was not always the case in the past.
- The set-theoretic model allows to reduce the complexity of the model through transitive closures when that much granularity is not required.
- The set-theoretic approach furthermore allows searches like "Give me a dictionary of a Dravidian language" without knowledge of all the languages contained within Dravidian. In traditional systems, one would have to gather all relevant ISO 639-3 codes by hand and iterate through them in individual queries.

6 Summary

Glottolog/Langdoc makes world-wide linguistics fit for the Semantic Web by providing near-complete bibliographical coverage of lesser known languages. It furthermore provides unique identifiers for its references as well as for languoids of every level, allowing third-party projects to harvest, digest, and further annotate the resources.

[8] `http://www.ethnologue.com/showlanguage.asp?code=lrk`

Bibliography

[1] Good, J., Hendryx-Parker, C.: Modeling contested categorization in linguistic databases. Proceedings of the EMELD 2006 Workshop on Digital Language Documentation: Tools and Standards: The State of the Art. Lansing, Michigan. June 20-22, 2006 http://www.linguistlist.org/emeld/workshop/2006/papers/GoodHendryxParker-Modelling.pdf (2006)

[2] Hammarström, H., Nordhoff, S.: Langdoc: Bibliographic infrastructure for linguistic typology. Oslo Studies in Language to appear, to appear (2011)