Modeling the relation between proto-languages and language families

Sebastian Nordhoff

June 28, 2012

Abstract

1 Introduction

language classifications Ethnologue Multitree Composite

(1) Indo-European
. Middle Indo-European
. Late Indo-European
. Germanic
. Northwest Germanic
. West Germanic
. North Sea Germanic
. Old English [ang]
. Anglian
. Mercian
. Middle English [enm]
. English [eng]

When interpreting the above tree, we can make a number of observations. We can say for instance that two adjacent nodes are in a childOf-relation. For instance, *Middle Indo-European* is a child of *Indo-European*, *Germanic* is a child of *Late Indo-European*, *West Germanic* is a child of *Northwest Germanic*, *Old English* is a child of *North Sea Germanic*, and so on.

A closer inspection, however, reveals that the nodes of this tree actually fall into two different sets. Some of them denote sets of languages, like *West Germanic*, others denote historic varieties, like *Old English*.

We can identify the sets by evaluating the proposition 'X is a subset of Y ' for adjacent nodes.

(2) a. *West Germanic* is a subset of *Northwest Germanic*
b. **Old English* is a subset of *West Germanic*
c. **Anglian* is a subset of *Old English*.

Historic varieties can be identified by evaluating the proposition 'X is a later variety of Y':

(3) a. *English* is a later variety of *Middle English*
b. **Middle English* is a later variety of *Mercian.*
c. **Anglian* is a later variety of *Old English.*

These are two important insights about the relation between languages and language families of Northwestern Europe at different points in time, but they should be kept distinct conceptually and the difference should be made explicit. At the same time, one would not want to completely dissociate the two, as there are obvious interconnections which can inform the historical linguist. How this conceptual separation can be achieved while maintaining the insights contained in the 'traditional' tree will be the topic of the remainder of this paper. I will show this on a stammbaum model as customary, which still seems to be the most popular model of representing language history. Future research will deal with the modeling of areality.

Both of the predicates just discussed are transitive:

(4) *English* is a later variety of *Old English*

(5) *Anglian* is a subset of *Indo-European*

Section **??** will show
Section **??** will show
Section **??** will show

2 How to define languages and language families: a resource based approach

The model proposed here is inscribed into a resource-based framework. Languages in the context of linguistic investigation are seen as a the collection of verifiable things we know about them. 'English' is the collection of everything ever written on English, 'Tocharian' is the collection of everything every written on Tocharian and so on. Note that there are some communication systems used by humans which will not be considered under this approach: those are the ones of which we have no documentation. This shortcoming is acceptable as these languages could not be classified anyway. While such languages certainly still exist on Earth, they have no bearing on issues of language classification while they are unknown. The second part of the definition is 'verifiable'. This excludes personal knowledge, personal communication, hearsay, conjectures etc. It must be possible for a dedicated researcher to trace the resources which form part of the definitional set used for a language. Note that the normal case for a resource is to be written, but audio or video resources are equally legitimate.

Within the set of 8031 languages in our database which have an official ISO 639-3 code, we have identified 6307, or 79%, which have at least one resource which treats them. We are still working on establishing the definitional set for the remainder. The amount of documentation available for the different languages is plotted in Figure 1

no fiat languages
Good, Nordhoff & Hammarström ISWC
problems with fiat languages

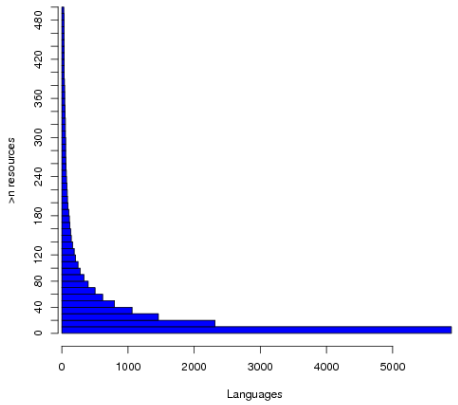

Figure 1: Number of languages which exceed a given number of resources. About 6000 languages have at least one resource; about 2200 of those have more than 10; about 1300 of those have more than 20 and so on. Only about 200 languages have more than 100 resources.

3 Definitions

We will handle the following concepts when modeling the relation between languages: *lectodocs*, *doculects*, and *languoids*. These will now be defined in turn.

3.1 Lectodoc

A lectodoc is a document which contains linguistically relevant information. The crucial point here is that we are dealing with documents, i.e. entities which could be printed. Lectodocs cannot be spoken. Dictionaries and grammatical descriptions are the prototypical lectodocs, but other kinds of work also fall under this definition. Word lists, theoretical articles on a particularly interesting phenomenon found in a language, or phonological descriptions are all documents which provide linguistic information and are therefore lectodocs. The types of documents just listed provide information about the phonological, morphosyntactic or lexical structure of a language. Next to these documents providing structural information, there are also lectodocs which provide information about sociodemographic details. These are equally important. Studies of language use, language variation, language and identity and other sociolinguistics topics come to mind. Related to this are ethnographic studies and census data. A final type of lectodoc are overview works like handbooks, bibliographies and the like.

3.2 Doculect

Cysouw.

A doculect is a documented linguistic variety (cf. dialect, idiolect, sociolect etc). It is the mental counterpart of a lectodoc. While a lectodoc can be printed, but not spoken, a doculect can be spoken, but not printed.

A doculect is tied to the lectodoc it is described in. This means that we have at least as many doculects as we have lectodocs. This definition entails that the OED and Webster's described different doculects. Actually, the granularity is even greater than that because different editions of the OED also describe different doculects.[1] To capture the intuition that these doculects are not so different after all, we can group them into languoids.

3.3 Languoid

A languoid is a collection of doculects. We can group the OED, Websters, and Quirk's grammar of English together and declare that these three doculects all provide information about a languoid we want to call 'English'. We have just formed a 'terminal' languoid, i.e. a languoid whose members are all doculects. It is also possible to have non-terminal languoids. Nonterminal languoids can contain other languoids. For instance, we could group a number of doculects into a languoid 'Dutch' and then group those two languoids together into a larger languoid 'West Germanic'. Nonterminal languoids may also contain doculects, e.g. 'The Germanic languages'.

Every languoid is reducible to the doculects contained in it. We call this list its *Bibliographical Grounding*.

4 Illustration

We will illustrate the three concepts just introduced with an example from Dravidian. We will start with the documents given in Table 1.

These seven lectodocs establish seven different doculects, which are grouped into two terminal languoids, Tamil and Malayalam in this case. The terminal languoids can then be grouped in turn into higher-order non-terminal languoids. The bibliographical grounding of the two terminal languoids and one higher-order languoid is given below:

- Tamil: Andronov 1960, Annamalai 2000, Asher 1985, Beythan 1943
- Malayalam: Andronov 1993, Asher 1968, Asher & Kumari 1997
- Tamil-Malayalam:Andronov 1960, Annamalai 2000, Asher 1985, Beythan 1943, Andronov 1993, Asher 1968, Asher & Kumari 1997

It is easy to see how this model would continue for higher order languoids like Southern Dravidian or Dravidian.

4.1 Basic modeling

concepts

The following relations hold between lectodocs, doculects, terminal languoids and non-terminal languoids:

- a lectodoc describes a doculect
- a doculect is an element of a languoid

[1] To put it differently, doculects correspond to FRBR 'manifestations'.

Lectodoc	Doculect	Terminal languoid	Non-terminal languoids
Andronov, Michail S. (1960) Tamil'skij jazyk	Tamil of Andrononov (1960)	Tamil	Tamil-Malayalam, ..., Southern Dravidian, ..., Dravidian
Annamalai, E. (2000) Lexical anaphors and pronouns in Tamil	Tamil of Annamalai (2000)	Tamil	Tamil-Malayalam, ..., Southern Dravidian, ..., Dravidian
Asher, R. E. (1985) Tamil	Tamil of Asher (1985)	Tamil	Tamil-Malayalam, ..., Southern Dravidian, ..., Dravidian
Beythan, Hermann (1943) Praktische Grammatik und Übungsbuch der Tamilsprache	Tamil of Beythan (1943)	Tamil	Tamil-Malayalam, ..., Southern Dravidian, ..., Dravidian
Andronov, Michail S. (1993) Jazyk malajalam	Malayalam of Andronov (1993)	Malayalam	Tamil-Malayalam, ..., Southern Dravidian, ..., Dravidian
Asher, R.E. (1968) Existential, possessive, locative and copulative sentences in Malayalam	Malayalam of Asher (1968)	Malayalam	Tamil-Malayalam, ..., Southern Dravidian, ..., Dravidian
Asher, R.E. and T.C. Kumari (1997) Malayalam	Malayalam of Asher & Kumari (1997)	Malayalam	Tamil-Malayalam, ..., Southern Dravidian, ..., Dravidian

Table 1: Lectodocs, doculects, and terminal languoids of some literature on Southern Dravidian.

Lectodoc	Doculect	Terminal languoid	Non-terminal languoids
Bloch, Jules (1954) The grammatical structure of Dravidian languages	Dravidian of Bloch 1954	-	Dravidian
Burrow, Thomas and Emeneau, M. B. (1998) A Dravidian etymological dictionary	Dravidian of Burrow & Emeneau	-	Dravidian
Caldwell, Robert (1998) A comparative grammar of the Dravidian or South Indian family of languages	Dravidian of Caldwell 1998	-	Dravidian
K. V. Zvelebil (1990) Dravidian Linguistics: An Introduction	Dravidian of Zvelebil 1990	-	Dravidian
Rangaswamy, R. (1995) Comparative Dravidian	Dravidian of Rangaswamy 1995	-	Dravidian

Table 2: Some works which directly attach to the non-terminal languoid 'Dravidian'

- a (terminal or nonterminal) languoid can be the subset of a nonterminal languoid
- a lectodoc forms part of the Bibliographical Grounding of a languoid

4.1.1 Terminal languoids

Terminal languoids have been modeled above and need no further discussion here.

4.1.2 Non-terminal languoids

Non-terminal languoids are distinguished from terminal languoids by their including other languoids. Doculects can be included, but this is optional. The following table lists lectodocs whose doculects would be directly included into the languoid 'Dravidian' rather than in any of its sublanguoids.

discussion

4.1.3 Historical varieties

We have now seen how the set-theoretic approach can model synchronic family relations between linguistic varieties and associated documents. When we include historical varieties, things become a bit more complicated. We can for instance not equate Tamil-Malayalam with Old Tamil. While it is perfectly

Lectodoc	Doculect	Terminal languoid	Non-terminal languoids
Lehmann, Thomas (1991) Grammatik des Alttamil: unter besonderer Berücksichtigung der Cankam-Texte des Dichters Kapilar	Old Tamil of Lehmann 1991	Old Tamil	Southern Dravidian, Dravidian
Steever, Sanford B. (2008) Old Tamil	Old Tamil of Steever 2008	Old Tamil	Southern Dravidian, Dravidian
Zvelebil, Kamil and Andronov, Michail S. and Navrozov, L. (1967) Vvedenie v istoriceskuju grammatiky tamilskogo jazyka	Old Tamil of Zvelebil, Andronov, & Navrozov 1967	Old Tamil	Southern Dravidian, Dravidian

Table 3: Some works which refer to the historical languoid 'Old Tamil'

legitimate to say that Asher & Kumari 1997 provide information about Tamil-Malayalam, it is wrong to say that they provide information about Old Tamil. Rather, we have to create a new terminal languoid 'Old Tamil', and model its relations to the extant 'Tamil', 'Malayalam', and 'Tamil-Malayalam'. This new languoid has to have Bibliographical Grounding as well. Table **??** provides some lectodocs for historic varieties.

On an non-formal level, we can formulate the following propositions:

- Tamil-Malayalam is the result of breakup of Old Tamil
- Tamil-Malayalam is not a later variety of Old Tamil
- Tamil is a subset of Tamil-Malayalam
- Malayalam is a subset of Tamil-Malayalam
- Tamil is a terminal languoid
- Malayalam is a terminal languoid
- Tamil is a later variety of Old Tamil
- Malayalam is a later variety of Old Tamil

Figure 2 illustrates this state of affairs. Striping denotes a historical languoid, the arrow symbolizes the 'breakup of' relation.

4.1.4 The definition of Proto-Dravidian

The above example has illustrated how to deal with a historical variety within a language family. In some cases, we have to assume that the languoid under discussion is outside the language family. This is the case for proto-languages

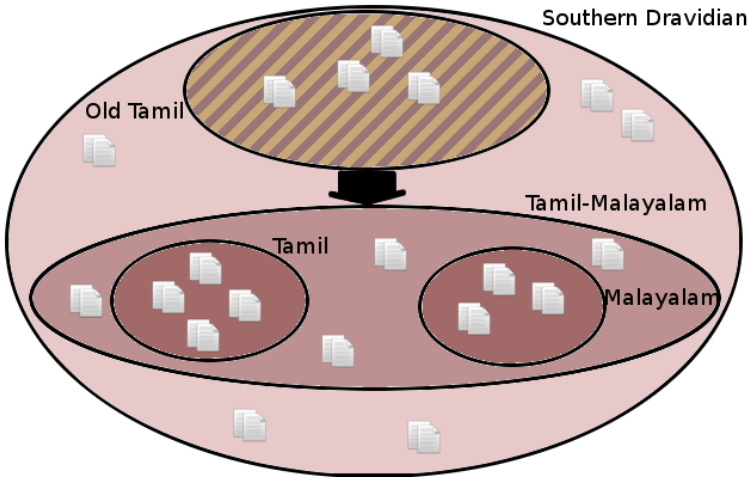

Figure 2: The relation between Old Tamil and Tamil-Malayalam. Old Tamil is a historical variety whose breakup is the non-terminal languoid Tamil-Malayalam, which contains the terminal languoids Tamil and Malayalam. Lectodocs can be associated with any of these varieties, either directly or indirectly.

of a whole family, e.g. Proto-Dravidian. These are by definition not part of the family whose latest common ancestor they are. Table 4 lists some lectodocs for Proto-Dravidian.

The breakup relation described above can also be applied to the relationship between Proto-Dravidian and Dravidian. Figure 3 illustrates this.

Theorems

(6) All terminal languoids found in the breakup set of a historical variety are later varieties of the historical variety

(7) No nonterminal languoid found in the breakup set of a historical variety is a later variety of the historical variety

Lectodoc	Doculect	Terminal languoid	Non-terminal languoids
William Bright (1986) Archaeology, linguistics, and ancient Dravidian	Proto-Dravidian of Bright 1986	Proto-Dravidian	-
Levi, Sylvain and Przyluski, Jean and Bloch, Jules (1929) Pre-Aryan and pre-Dravidian in India	Proto-Dravidian of Levi, Przyluski & Bloch 1929	Proto-Dravidian	-

Table 4: Some works which refer to the historical languoid 'Proto-Dravidian'

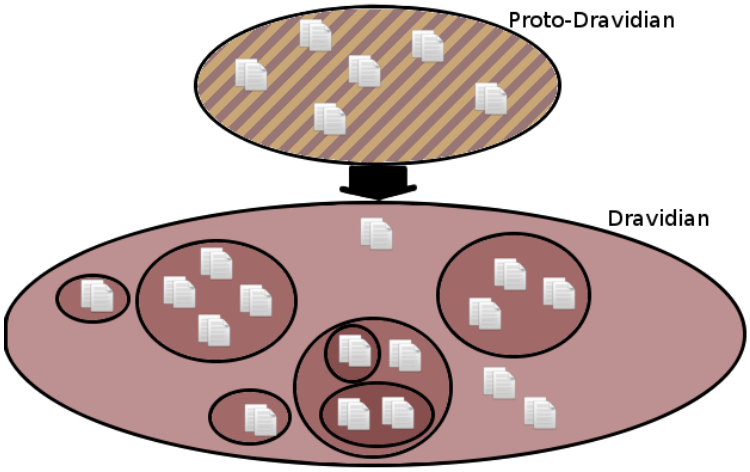

Figure 3: A top-level family as the result of the breakup of a proto-language which his not a member of a family itself. This case is illustrated by Proto-Dravidian.

4.2 A note on dialects

Note that not everything which is currently seen as a language automatically corresponds to a terminal languoid. This is most notable the case for the so-called macro-languages, like Arabic, Kurdish, or Quechua. These macro-languages are important at a social level, and speakers often identify with them, claiming to be a member of this linguistic group. At the same time, mutual intelligibility is often low, and the various varieties of Arabic for instance can be considered languages in their own right. This means that 'Arabic' is not a terminal languoid, but a nonterminal languoid, as it subsumes a number of subvarieties. The relation between the Arabic dialects and the macro-language Arabic is thus the same (in a set-theoretic sense) as the relation between the macro-language Arabic and Semitic.

The same hold mutatis mutandis for dialects of English. There are abundant descriptions of English dialects, which can be grouped into non-terminal dialects. The languoid 'English' then contains all those subsets, and is therefore nonterminal. Note that 'English' is used here as a label for a set of varieties, and is distinct from 'Standard English', which *is* a terminal languoid. This terminal languoid 'Standard English' is itself a subset of the non-terminal languoid 'English'.[2]

4.3 A note on historic dialects

There are cases where we have a succession of historical varieties without necessarily a break-up stage in between. Archaic Latin and Classical Latin are for instance normally seen as continuations of the same language without intervening speciation. The model discussed here does not allow for such direct

[2]In order to avoid confusion, one could choose 'Angloid' as a label instead of 'English' to highlight the set nature of the concept.

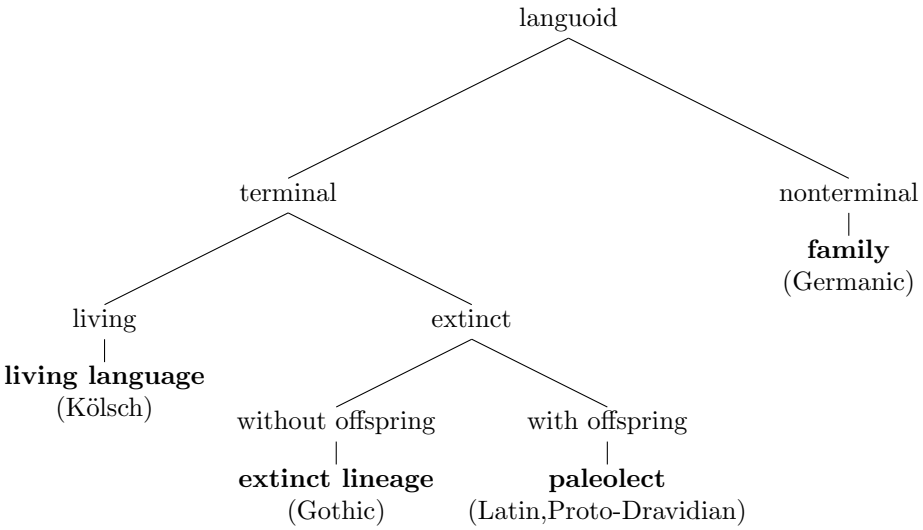

Figure 4: Hierarchy of languoids with types and examples.

	living	extinct lineage	paleolect	family
contains other languoids	–	–	–	+
has speakers	+	–	–	+/–
has scatter	–	–	+	–

Table 5: Properties of languoids

successions. Instead, it assumes a breakup of Archaic Latin into a very small family consisting only of Classical Latin, a singleton family so to speak. This has the advantage that the model does not have to be changed if a sister of Classical Latin were to be found. Hard-coding of the succession relationship between Archaic Latin and Classical Latin would require remodeling upon the discovery of such new knowledge, which is undesirable from an information science point of view.

5 Types of languoids

After this basic introduction, we can now engage in a more formal representation of the different types of languoids. Table **??** shows the hierarchy of languoid types.

Basing ourselves on the criteria [±terminal], [±extinct],[±offspring], we can establish four types of languoid: **Families** are non-terminal languoids. Terminal languoids are divided into **living languages**, which do have speakers, and historic varieties, which do not have speakers. Historic varieties are in turn divided into **extinct lineages** without any offspring and **paleolects**, which are the ancestor of at least one other terminal languoid.

These four types of languoids can be dinstinguished by a number of properties, which are listed in table 5.

Only a family can contain other languoids, a logical corollary of the other

	living	extinct lineage	paleolect	family
living	/	/	succession	inclusion
extinct lineage	/	/	succession	inclusion
paleolect	/	/	succession	inclusion
family	/	/	scatter	inclusion

Table 6: Relations between languoids. The 'lower' varieties are on the left and the 'higher' varieties are on top.

languoids being terminal. The prime example of a languoid with speakers is the living language. A family can also have speakers if it contains at least one living language. If it contains only paleolects and extinct lineages as languoids, it does not have any speakers. The third and final concept is 'scatter'. This refers to the breakup of a paleolect into a new language family. It is thus the projection of a terminal languoid on a non-terminal languoid, as discussed above for Old Tamil and Proto-Dravidian.

These three criteria provide operational definitions for the classifications of languoids into the four types.

These types themselves can enter in relations with each other as shown in table 6

We see that living languages and extinct lineages cannot be found on the 'older' part of the relations. What is found on the 'older' part are paleolects and families. Families can inlcude any of the other types of languoids. The relation between paleolects on the one hand and living languages or extinct lineages is one of succession. The relation between paleolects and families is called 'scatter' here, i.e. the result of breakup. We thus have to model three relations: inclusion, scatter, and succession. Additionally, we have to get back to the doculect/lectodoc issue and find a formal modeling for this as well.

6 Mathematical modeling

In the following, we will use L to denote a languoid, p to denote a paleolect, dl to denote a doculect and ld to denote a lectodoc. Subscript t and nt refer to terminal and nonterminal languoids, respectively.

6.1 Inclusion

$L_t \subset L_{nt}$

$L_{nt} \subset L_{nt}$

6.2 Scatter

$p \prec L_{nt} \prec$

6.3 Succession

$p \prec L_{nt}; L_t \subset L_{nt}$

6.4 Doculects

$dl \in L$

6.5 Lectodocs

$ld \hat{=} dl$ *lddl* every lectodoc describes at least one doculect, potentially more *dlld* every doculect is described by exactly one lectodoc

7 Ontological modeling in the Semantic Web

7.1 Overview

intro OWL concepts instances

7.2 inclusion

skos:broader skos:narrower
reflexive + transitive (transitivebroader) + symmetric -

7.3 elementof

rdf:a
reflexive - transitive - symmetric -

7.4 precedence

$\prec$ reflexive - transitive - symmetric -

7.5 equivalence

$\hat{=}$
powder:describes
reflexive - transitive - symmetric -

7.6 Succession

$p \prec L_{nt}; L_t \subset L_{nt}$ reasoner
reflexive - transitive + symmetric -

7.7 Other reasoners

Latest common ancestor Empirical grounding

8 Glottolog

9 conclusion and outlook

outlook extension to areal linguistics