DiACL – Diachronic Atlas of Comparative Linguistics Online. Database description

Author: Gerd Carling (Lund University)

Date of publication: 2017-03-08

Contributors (dataset design, feeding, language control/ contribution, database construction, programming, advice): Gerd Carling, Filip Larsson, Sandra Cronhamn, Rob Farren, Niklas Johansson, Anne Goergens, Rob Verhoeven, Arthur Holmer, Chundra Cathcart, Erich Round, Karina Vamling, Maka Tetraze, Tamuna Lomadze, Merab Chukhua, Leila Avidzba, Teimuraz, Gvadzeladze, Madzhid Khalilov, Ketevan Gurchiani, Revaz Tchantouria, Ana Suelly Arruda Câmara Cabral, Joaquim Mana, Vera da Silva Sinha, Wany Sampaio, Chris Sinha, Nanblá Gakran, Wary Kamaiurá Sabino, Aisanain Páltu Kamaiura, Kaman Nahukua, Makaulaka Mehinako, Johan Dahl, Edin Kukovic, Johan Åhlfeldt.

Reference, database: Carling, Gerd (ed.) 2017. Diachronic Atlas of Comparative Linguistics Online. Lund University. Accessed on: z

URL: https://diacl.ht.lu.se/

Reference, document: Gerd Carling (2017). DiACL – Diachronic Atlas of Comparative Linguistics Online. Database description. Lund University, Centre for Languages and Literature.

Table of Content

§1. The DiACL database

§1.1. Background and aim

The aim of the DiACL database is to provide users with an open access database resource for reconstructing language evolution and change in history and prehistory. The focus of the database is diachronic and laid on filling data sets as far as possible synchronically and diachronically, using predefined, selected feature sets, which can be extracted and analyzed by means of computational methods. The database aims at providing, as far as possible considering the constraints provided by data access, data from historical and reconstructed languages. This basic aim is an important prerequisite both for the selection of targeted features, and the database design. The targeted data in the database is close or corresponding to data types that are frequently used in other databases for the purpose of computational cladistics research, i.e., lexical data (basic and culture vocabulary), and typological data (word order, alignment, nominal/verbal morphology). The database design and the character coding has the following aims: 1) the targeted data sets have a high granularity in character coding, 2) data sets are pre-prepared for computational analysis, which means that they are filled to a satisfactory level (and need not to be filtered), 3) data sets contain historical and reconstructed data, as far as possible, implementing comparative method.

§1.2. Rationale

The rationale of the DiACL database follows a principle of providing data sets which reflect, with a high degree of granularity and accuracy, retained wisdom from comparative-historical linguistics. The underlying aim is to adapt comparative-historical linguistics to computational methods of analyzing data by means of a method of data character coding that encapsulates historical-comparative information. This concern, e.g., geographic position and temporal extent in prehistory, or known prehistoric language contact, e.g., by lexical borrowing.

The theoretical model, underlying the rationale of the database, is a digitization of the space-time model (*Raum-Zeit-Modell*) within historical-comparative linguistics (Meid, 1975), which refers to a stratified model, which fixes languages and linguistic patterns in time and space, based on a grid of language documentation, contemporary and historical, as well as patterns reconstructed by relative chronology and internal/external reconstruction. With an increasing time-depth, the basis for a reconstruction in time and space shrinks. In implementing the model, the database uses *language* (not dialect) as a unique identifier, and this concept includes contemporary, historical, as well as reconstructed languages. By necessity, *languages* are definable units, which can be constrained into time and space, but with a varying degree of certainty. Therefore, the unique identifier *language* is connected to metadata information that includes time period, spatial extent (focal points/ polygons), position in a cladistic reference tree, and reliability (modern language/ dead language (well documented)/ dead language (fragmentary)/ reconstructed language). All this information is retrieved by conventional methods and are sourced in scientific literature.

The unique identifier *language* links to tables with linguistic data. Feature organization and data character coding is implemented according to a hierarchical model of increasing granularity, which is implemented both in lexical and typological data. This hierarchical model aims at capturing dynamics at the functional side of the data, enabling definitions and partitions of data sets for testing various models in data analysis.

From the database, XML or JSON-files of distinct datasets, pre-prepared for statistical analysis, can be extracted. These extracted files reflect the internal hierarchical structure of the database design and carry along all of the information available in the database. These extracted sets can be used for statistical testing and reconstruction.

§1.3. Language areas and languages

At current state, the data of the DiACL database embraces three macro-areas, defined as *Focus areas*: Eurasia, Amazonia, and Pacific. Datasets are available from the families of Aikanã, Arawakan, Austronesian, Basque, Cariban, Chapacuran, Indo-European, Jabutian, Jêan, Kanoé, Kartvelian, Nambikwaran, North-East Caucasian, North-West Caucasian, Pano-Takanan, Tupían, Turkic, Uralic (see fig. 1). The most important family, with the largest data coverage and granularity, is the Indo-European family of Eurasia, since Indo-European has the richest availability and reliability when it comes to comparative linguistic and historical sources. The Indo-European family has served as a model when the principles of database design and data character coding have been developed. For the Amazonian and Pacific areas, data sets contain fewer languages and more restricted number of features, and naturally, no historically attested languages are available. Other families, such as Uralic, and Turkic, have currently restricted number of languages, but complete feature sets. All over, languages are organized into phylogenetic tree topologies, and reconstructed states at various levels (e.g., Proto-Tupí, Proto-Arawak, Proto-Indo-European, Proto-Slavic) are defined and filled with reconstructed forms, if available or else reconstructable.

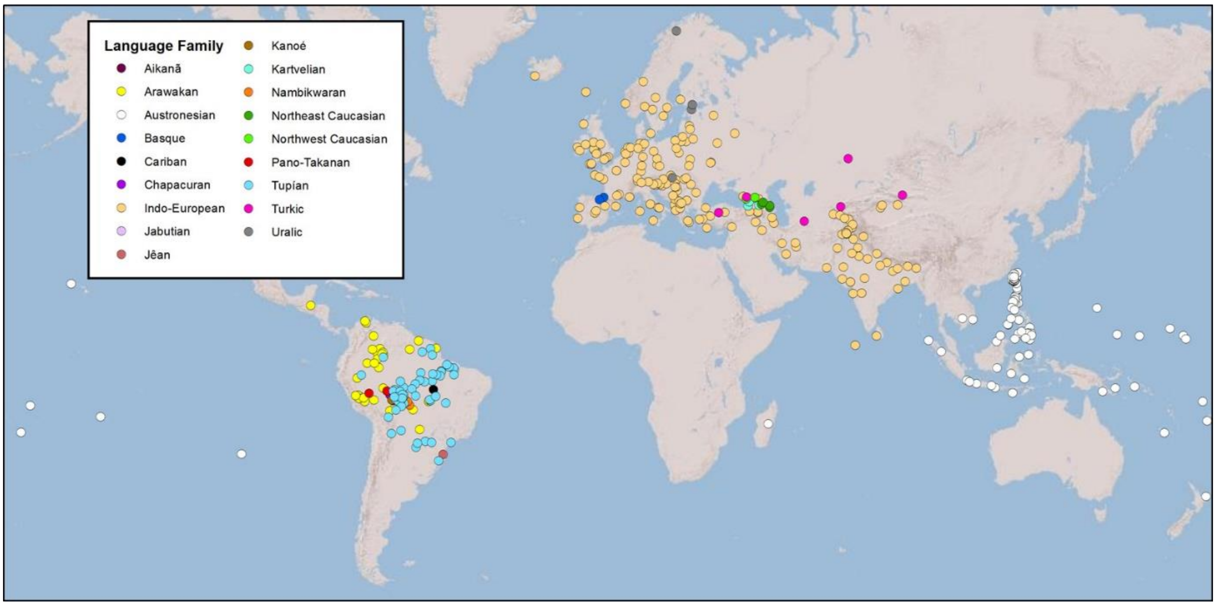


Figure 1. Overview of languages, organized into families, with data in the database

§1.4. Data types and design policies

The DiACL database contains four main types of data, representing four subsections of the database (fig. 2):

1) Language metadata (including geographic position, temporal extension, reliability, and family tree topology),
2) Lexical and etymological data, including Swadesh-lists and culture vocabularies,
3) Typological/morpho-syntactic data, including word order, alignment, nominal and verbal morphology,
4) Source data.

Lexical and typological data are frequently used in computational cladistics (Nichols & Warnow, 2008) and therefore also in focus here. Language metadata on spatial and temporal extension and family tree topology are useful for all types of computational modelling. Every datapoint in the database, including geographic data, is supposed to be sourced in scientifically reliable literature.

The model of the DiACL database is to organize data sets into functionally hierarchical categories, the purpose of which is as follows:

1) To level out language-internal possible polymorphic behavior and create informative, dynamic, and representative datasets of high granularity,
2) To enable partition of data sets on the functional side, to test language-internal differences systematically,
3) To create data sets that are representative, symmetrical, and with a high degree of completeness.

The basic model of the database, both for typological and lexical data, aims at a hierarchical organization at the functional side, where the top-most level (below the level of Focus areas) is more general, "universal", whereas lower levels are adapted to an observed reality of Focus areas. Currently, three geographic Focus Areas are defined: Eurasia, South America, and the Pacific. These three areas have different data sets for both lexical and typological data, with the exception of basic vocabulary (Swadesh list), which is the same for all languages and therefore not distinguished by Focus area. Even though the overarching principle of adaptation is similar between typological and lexical data sets, the outcome of the feature and character selection and organization is very different, due to the fundamentally different nature of typological and lexical data. Likewise, there are differences between data sets of different Focus areas, due to the different nature of linguistic realities of different areas. However, the overarching principle for compiling data sets is similar for all Focus areas, and for that purpose, the aim is that data sets should correspond to each other at a more general level.

§3. Database structure: tables and relations

The core of the database is the entity *Language*, which contains languages along with attributes, to which all other sections of the database link. Language includes contemporary languages, extinct languages, and reconstructed language states (e.g., Proto-Indo-European, Proto-Tupí). Four tables constitute the extended core of the database (fig. 2):

Tables:	**Targets:**
1. Language	Language and language metadata
2. Language Tree	Position in family tree topology
3. Focus Area	Geographic macro-area
4. Geographical Presence	Geographical presence (focal point/ polygon)

The table Language contains the metadata information Focal point (point on a map), Polygon, Area, Focus-area, Reliability, and Time Frame.

Outside of the core, the database has three subsections: 1) Lexicology, 2) Typology, and 3) Source. The organization of the subsections Typology and Lexicology are described in other documents.

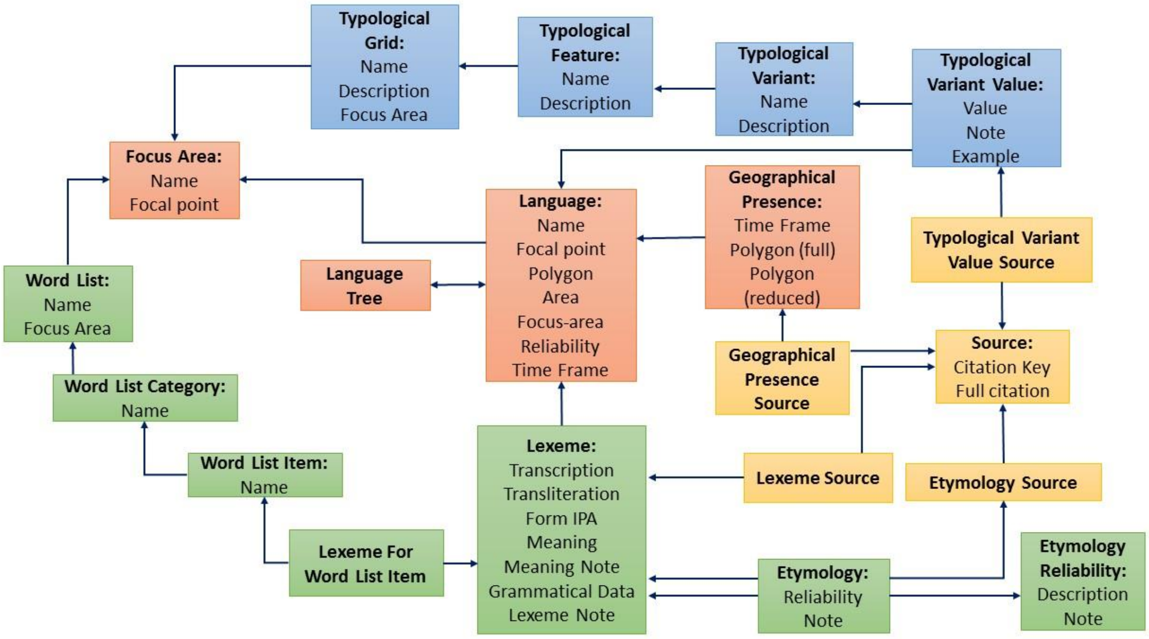


Figure 2. Tables and relations in the database. Orange: Language and language metadata, Blue: Typology/morphosyntax, Yellow: Source, Green: Lexicon.

§2. Language metadata: time, space and family tree topology

Language metadata on the table Language (fig. 2) includes a standardized name, ISO 693-3 code, alternative names, location, time frame, language area, and reliability (fig. 3).

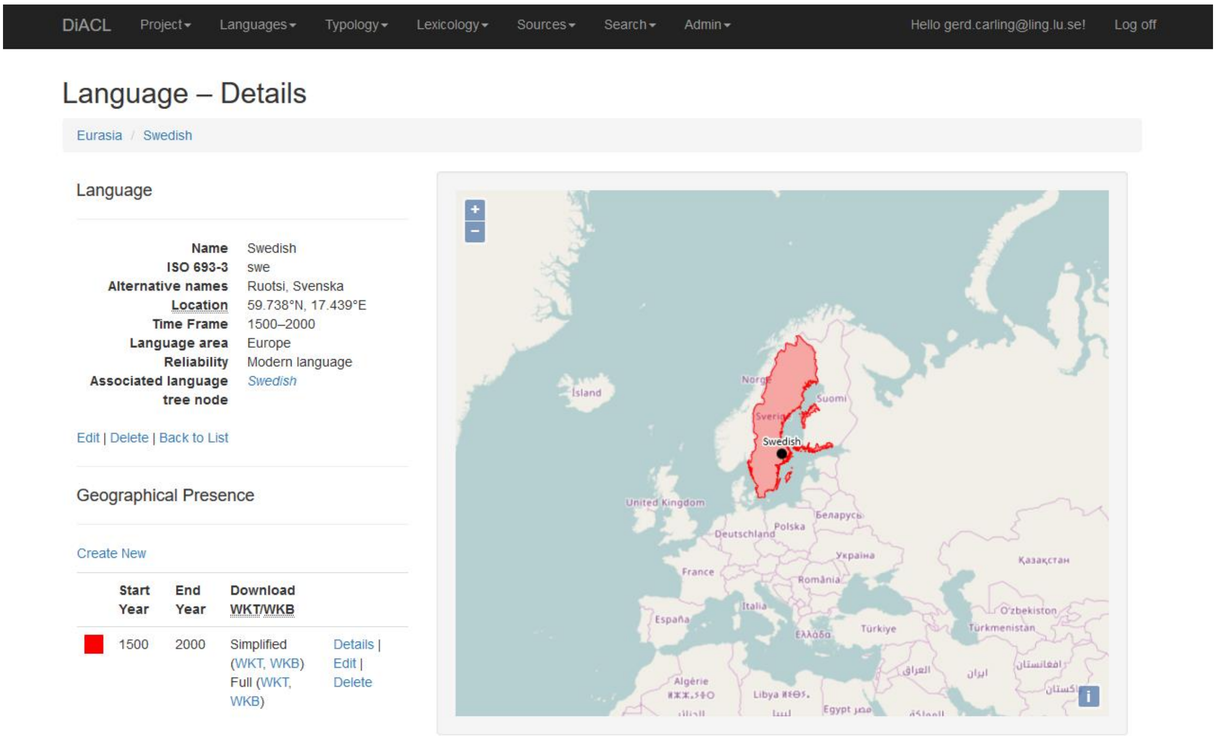
Location gives a focal point, which renders the most proto-typical geographic center for a language. In the database, the focal point typically is centered within the area where the standard variety of a language is spoken. For some languages, areas are slightly adjusted for avoiding of overlapping when languages of different time periods are viewed on maps.

Alternative names gives various names of the language in different descriptions or in different languages.

Time frame gives an estimation of the period within which a language is spoken. These dates are not founded in specific historical events: they are rather estimations, given in 100-year intervals.

Language area is a more detailed classification of language areas, compared to Focus area (see §1.3).

Reliability has four distinctions: Modern language, Dead (well attested), Dead (fragmentary), and Reconstructed. Dead (well attested) targets languages with corpora large enough to provide data equivalent to living languages, whereas Dead (fragmentary) targets languages with lesser documentation: for these languages, data sets have to be controlled for completeness. Reconstructed targets languages that have no literary sources: they are entirely based on reconstruction by comparative method. Reconstructed languages contain no typological/ morphosyntactic data. They may contain lexical data, but always provided with an *, marking that the forms are reconstructed, not attested.



Language – Details

Eurasia / Swedish

Language

Name	Swedish
ISO 693-3	swe
Alternative names	Ruotsi, Svenska
Location	59.738°N, 17.439°E
Time Frame	1500–2000
Language area	Europe
Reliability	Modern language
Associated language tree node	*Swedish*

Edit | Delete | Back to List

Geographical Presence

Create New

	Start Year	End Year	Download WKT/WKB	
🟥	1500	2000	Simplified (WKT, WKB) Full (WKT, WKB)	Details | Edit | Delete

Figure 3. Screenshot of Language Swedish, with language metadata

Focus area is a main distinction, separating subsets of data in the database. Languages are defined by Focus area, of which there are currently three: Eurasia, South America, and Pacific (§1.3). These Focus areas are macro-areas, to which data sets are adapted.

Languages are also classified by Language Tree, which defines the position of languages in a (handcrafted) tree topology. In these trees, proto-languages, which have their own ID and metadata, are located at nodes within the trees.

§4. Technological implementation, sustainability

The database DiACL (https://diacl.ht.lu.se/) resides in Microsoft SQL Server 2014, making use of several of its specialized data types for recording hierarchical and geographical information. Its online interface has been made in ASP.NET MVC 5 (which on the server side employs the model-view-controller architecture incorporating the *repository* and *unit-of-work* patterns). On the client side, the interface makes use of OpenLayers 3 to display maps and the jQuery library for added responsiveness.

Languag area (multi)polygons wore originally drawn in ArcGIS, exported to Well-Known Text (WKT) with Python script (in desired projection WGS84), which can be inserted in SQL Server database using the geometry datatype. Via the interface, it is possible to download WKT and WKB files with language area multipolygon data.

Both the database and the online interface reside on an IIS server currently hosted by the Faculties of Humanities and Theology at Lund University. In the future, sustainability of the database will be secured through SWE-CLARIN at Lund University (https://sweclarin.se/swe/centrum/lund), an initiative by **ESFRI (**http://www.esfri.eu/**).** DiACL is also a SND resource (https://snd.gu.se/sv).

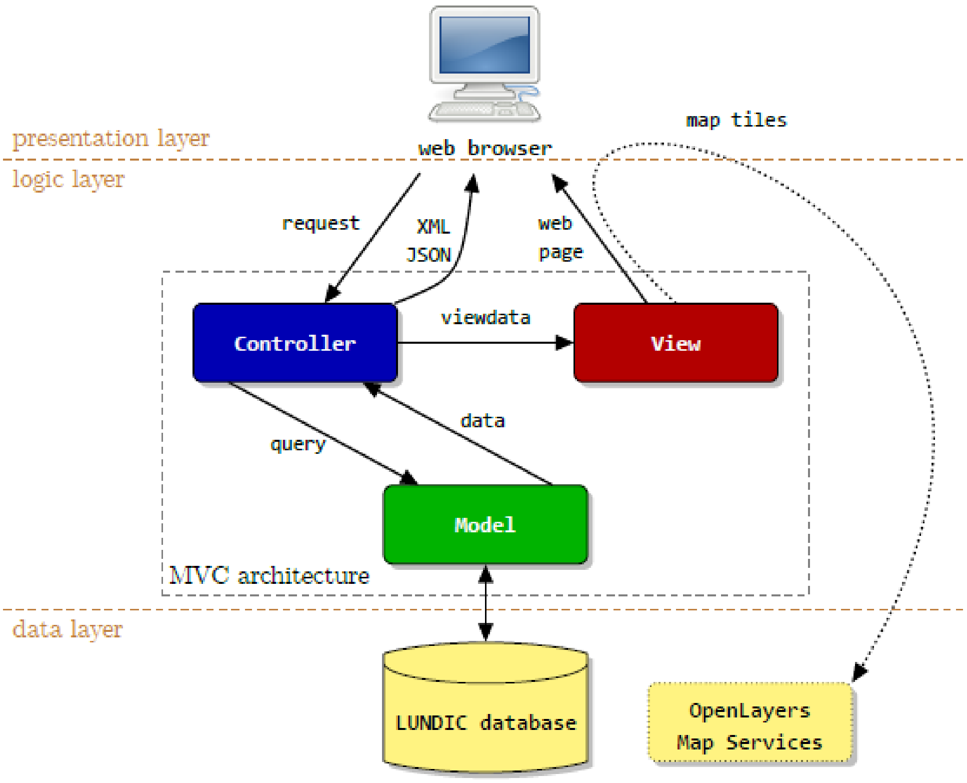


Figure 4. Technical sketch of the database.

Meid, W. (1975). Probleme der räumlichen und zeitlichen Gliederung des Indogermanischen. In H. Rix (Ed.), *Flexion und Wortbildung* (pp. 204-219). Wiesbaden: Ludwig Reichert.

Nichols, J., & Warnow, T. (2008). Tutorial on computational linguistic phylogeny. *Language and Linguistics Compass*, 2, 760-820.