Conlang Database Project

Introduction

A group of us have initiated a project to create and maintain a database of the world's conlangs. This database will aim to collect as many conlangs as possible in one searchable catalogue, with useful information about them.

Conlanging has increased exponentially in recent decades. There are many lists of conlangs, for instance in the pages of prominent conlangers, or in the few books that have been written about the subject; however there is no comprehensive catalogue or database that gathers information about as many of them as possible in one place, for the benefit of anyone interested in knowing about them. There are resources that aim to feature many conlangs, such as the wikis like FrathWiki and Linguifex; or CALS, the Conlang Atlas of Linguistic Structures; or ConWorkShop which is a one-stop shop resource for conlangers. However these do not try to collect all conlangs simply because they're conlangs, and that's what we want the database to do.

It has been claimed that at this point it's extremely unlikely that we could record all conlangs, or even all those of which there is findable evidence on the internet. There are simply too many now, and more and more are being created all the time; finding them and adding their information to a database would take too much time to catch up. This may or may not be true, but even if true it's not a reason not to have the database. Instead the database would simply aim to record more and more languages over time; most of the more notable ones would be included quite soon, and there would simply be more and more other ones added thereafter. It's also possible that we could add all findable conlangs, if we had enough volunteers, with enough work time, and organised well enough with division of labour.

Online project locations

Currently this database project has the following locations online:

Discord channel: https://discord.com/channels/748166870787555448/
(invite link: https://discord.gg/J55RyKd)
Facebook group: https://www.facebook.com/groups/2757660151170529
Reddit subreddit: https://www.reddit.com/r/ConlangDatabase/
Planning GDoc:
https://docs.google.com/document/d/1y5QhCZpn3XcaosbP3wAEtoNGluLHHLzvateNZNAI6Ek/
Data Input Spreadsheet:
https://docs.google.com/spreadsheets/d/1uulyGvrhxHxqiY9LnQ8e5hGbENzmEqTwEr0qeyWtsSo/

There are several other planning documents for subsections of the plans:
New proposal for fields and options:
https://docs.google.com/document/d/1yrqSWOq1VTcpzvCxHlfI-MjKxLMZC2m6VNAejfmZgH4/
Modus Operandi:
https://docs.google.com/document/d/1IDhVDNEtXjoFv1W-hnrg6O6UqHVsChSzMVwmSFsC66k/

Principles

The project as it stands is shaped by certain principles and intentions:
- The database is not meant to be private, but publicly viewable and searchable.
- It should be open to any constructed language, even at a minimal level of development.
- It is only intended to include public domain information, not copyright material such as the details of languages that creators have written.
- It should be based in one location, but data should be backed up in other places to secure it against future loss of its specific host.
- It should be searchable for specific character string content (e.g. to find a conlang of a particular name) but also for lists of all conlangs of particular types or combinations thereof.
- The work of adding entries should be done by an organised and overseen team of volunteers working in co-ordination.
- There are many different possible sources of conlang data and different volunteers could concentrate on different ones at the same time – they might also have different priorities which is fine as long as there is a coherent overall policy.
- Besides volunteer work building the database, there should be a website input form for people to add conlangs, such as their own or ones they know about – this would probably greatly speed up its growth.
- There are likely to be disagreements over some content such as classifications of conlangs; these should be resolved as amicably as possible with a mind to both honour creators' feelings and wishes and maintain accuracy in the database.
- If a creator really doesn't want their language included in the database, it should be OK to remove it... however all languages featured would be in the public domain so there shouldn't be any legal or moral issue around including them.

Plan for implementation

The Language Creation Society (https://conlang.org/) have agreed to host the database, on the subdomain "database.conlang.org". We will be using MySQL for the database and PHP for the search forms. We will be able to back up the data in various locations, in order to avoid the problem of the data being lost if an individual website or data store is lost. We will use a "CC-BY-SA" licence (https://creativecommons.org/licenses/by-sa/2.0/) which means the data will be usable by all for free.

The website will have:
1. an explanation of what it is about, perhaps with some examples
2. a search form for people browsing to search for conlangs
3. an input form for them to enter details of conlangs
4. a contact form for people to get in touch
5. possible optional extras such as articles, links, a directory

The project would have the following roles:
Contributors (members of the public not of the project) will submit conlangs via the public input form on the website. **Reviewers** will look over public contributions and check them for appropriate content. **Compilers** will actively go out to search for as-yet un-added

conlangs and add them direct to the database. **Organisers** will have an executive role, setting policy and direction and co-ordinating the others.

Searching the database will involve a combination of whichever and however many fields are needed, with wildcards where possible as well as general keywords. The results will display according to the searcher's choice and may be printed or downloaded. Contributing conlangs will involve the same fields for entering data. Contributed conlangs will not be put on the publicly viewable database until checked by the reviewers. Messages entered via the contact form will have optional message type settings (question, suggestion, complaint) and go to the organisers for a response.

Reviewers and compilers will have some level of admin access to the database 'behind the scenes' in order to check public-contributed data, or to add languages more quickly or in more versatile ways than using the public form. Organisers will oversee this work including keeping lists of conlang sources and assigning them to compilers. They will also direct changes to the website or database and represent the project to the world.

Further details can be found in the 'Modus Operandi' document (see above in Online Locations).

Data fields

1. Initial setup

A couple of us created a spreadsheet to demonstrate the database idea and as a start to gathering conlang data - see above for the link. It contains the following fields:

Name	Name of conlang in English/interlingual version
Endonym	Name of the conlang in itself
Creator	Conlanger's name. In the case of languages created by a group, the name of the group.
Page	Clickable link to the most important page/site for the language
Start Year	The year the creator started creating the language
Mode(s)	What physical media the language is expressed in, such as speech, writing, signing, other rarer media.
Purpose Type	Classification of the conlang by the purpose for its creation – at the most basic level, e.g. auxiliary, artistic, engineered
Purpose Subtype	Sub-classification by purpose, at the next level – e.g. fantasy, alternate history, logical
Source Type	Classification of the conlang by the origin of its content (especially vocabulary) - whether it's 'a priori' (made up from scratch), 'a posteriori' (derived from previous languages), or a mixture.
Lexico-Semantics	Classification of the conlang according to the organisation of its vocabulary in relation to areas of meaning. E.g. taxonomic vs. naturalistic

Development Level	How developed the language is, from sketchlang at the minimum to native speakers at the other.

These fields are of two types:
1. Basic identifying information of the language, such as name, name of creator, date of creation etc. These are generally factual (right or wrong) and should be public domain information. The database input style should be open, although the **Date** should of course be a number field etc.
2. Classification categories for the language, such as purpose or source types, development level, physical modes used. These are more a matter of judgement or classificatory choice, and are not necessarily explicitly stated by the creator or in documentation (but it is better if they are). It might be best if the database input style for these fields is a selection from a limited list, or else contributors might try all sorts of random entries and compilers will only have to change it later.

The above data fields had the following settings for possible entries:

Name	Text string
Endonym	Text string, possibly with non-standard characters. May include an image of the conlang name in its own conscript.
Creator	Text string
Page	Clickable URL link, possibly several (potential multiple lines)
Start Year	Year number
Mode(s)	Options from menu: "Phonic-Scriptic" (presented as spoken and then written) "Scriptic (+Phonic)" (presented as written, with the implication that it is or could also be spoken) "Phonic only" (spoken without written version; only written in IPA) "Pasigraphic" (written language directly representing concept rather than speech) "Sign" or "Chiric" (manual sign language) "Musical" "Other"
Purpose Type	Options from menu: "Artistic" (artlangs, for aesthetic purposes) "Auxiliary" (auxlangs, to help language communities) "Engineered" (engelangs, to test theories, principles and designs) "Personal" (created only for the creator's own enjoyment) "Jokelang" (created purely for amusement) "Mystical" (inspired mystically or used for ritual purposes) "Secret" (stealthlangs, for clandestine communication) "Other"
Purpose Subtype	Options from menu: Artistic: ● "Fictional" (created to feature in a work of fiction) ● "Conworld" (created as part of imaginary (constructed) world)

	● "Geofictional" (imagined as being in a fictional part of our world)
	    ● "Alternate History" (imagined as part of alternative timeline)
	    ● "Lostlang" (imagined as undiscovered part of our timeline)
	    ● "Other"
	Auxiliary:
	    ● "Global" (intended for use by whole world)
	    ● "Zonal" (intended for use by limited area of the world)
	    ● "Specialist" (intended for use by specific section of humanity)
	    ● "Other"
	Engineered:
	    ● "Ideal" (created to attempt an optimum representation of ideas)
	    ● "Philosophical" (expressing a philosophical viewpoint)
	    ● "Logical" (aiming to express meaning without ambiguity)
	    ● "Experimental" (testing or demonstrating linguistic possibilities)
	    ● "Other"
	"Other Type" (setting for the other purpose types besides the main three)

Source Type	Options from menu:
	"A posteriori" (derived from other languages)
	"A priori" (not derived from other languages)
	"A priori" could be divided into:
	    ● "Randomly generated" (e.g. using a word gen script)
	    ● "Personally generated" (as in the creator choosing everything individually)
	"Randomly generated" itself could be divided into:
	    ● "Randomly assigned"
	    ● "Personally assigned" (if the creator generates words randomly and then assigns meanings by choice)
	"Mixed" (a mixture of derived and not derived from other languages)
	"Other"

Lexico-Semantics	Options from menu: "Taxonomic" "Naturalistic" "Other" (maybe there are other types to be named?)
Development Level	Options from menu: "Unknown" "Beginning" or "Incipient" (minimal progress, either a sketchlang or superficial fictional language) "Developed" (the language has been built to a usable level) "Learners" (the language has attracted people to learn it besides its creator) "Active Community" (users are conversing and developing the language) "Fluent Users" (some learners have achieved fluency) "Native Users" (some children have learned the language as a mother tongue) "Other"

2. Revision by the group

The above set of fields was OK for an initial version, but isn't necessarily the best for the database. It would undoubtedly be better if we could finalise the fields before we start the database proper, as it would require a huge amount of reorganisation and extra work if we changed them after adding lots of conlangs. From the start I felt that there should be a discussion amongst all those interested in creating the database as to how they could be improved. Also that time should be spent thinking about it rather than rushing into a possibly inadequate solution.

One factor is: if there are too many fields it might put people off from bothering to enter data at all. It is likely that many fields will have to be open to being left empty, as the contributor may not know or care about the answer, or there might not be a clear answer. If compilers could fill these in later it would be good, but it may be the case that some are left empty permanently. What fields are chosen and how many is a balance between including as much interesting info on the one hand and keeping the scale of input down on the other. It might be a good idea if we make it clear that filling in all fields is not compulsory, guiding contributors to concentrate on the most important/relevant factors.

There have been the following recommendations from different people as to changes to the fields:

- Add a general **Notes** field at the end for creators or compilers to add noteworthy information that wouldn't be contained.
- Remove **Lexico-Semantics** – it is only really there to identify taxonomic languages, that info can go in **Notes**.
- Collapse **Purpose Type** and **Purpose Subtype** into one. The latter is determined partly by the former anyway, so the subtype categories can just have the type ones added to the beginning, e.g. "Artistic – Fantasy".
- Make sure the **Page** field can contain several links, or else have a second field for extra links.
- Add a field for **Pseudonyms** of creators. This would be useful when someone only knows a pseudonym and not a real name.
- Add a **Group** field for conlangs in either a natlang or conworld genealogical grouping. Could also be used for different languages in the same suite.
- Add a field for what **Scripts** the language uses.
- Add a field for **Previous names** of the conlang.
- The **Endonym** field could include an option to upload an image of the language's name in its own conscript.
- Add a field for Conlang **Code** Registry (CLCR – https://www.kreativekorp.com/clcr/)
- Add a "**Feature Summary**" field for notable features of the language - or else make clear that the **Notes** field can contain "features of interest". These features are not meant to be the WALS-type features of CALS, but more intriguing and peculiar qualities of the language.

If the above data fields were adopted as suggested they would have the following settings for possible entries:

Notes	Open content (potential multiple lines)
Pseudonym(s)	Text string (potential multiple lines)
Group	Text string
Scripts	Text string (potential multiple lines)

Previous names	Text string (potential multiple lines)
Code	3-letter text
Feature Summary	Text string (multiple lines)

3. New proposal and finalisation Dec 2020

I've taken on board a lot of ideas and have tried to cover all reasonable issues in the new scheme below.

Name(s) of Language (in English/interlingual)

Should be multi-line to allow for more than one name. Can include former names.

Name(s) of Language (in itself)

Should be multi-line to allow for more than one name. Should also have a facility to allow for upload of an image file showing the name in the language's conscript.

Conlang Code Registry code

Gives the code (up to 8 letters long) the CLCR has applied to the language, if it has.

Name(s) of Creator(s)

Should be multi-line to allow for more than one name per creator, and also for more than one creator. Neither real names nor pseudonyms should be required, it should be up to the creator which or what they prefer. Real names should be entered such that they can be ordered alphabetically by family name. In the case of languages created by a group, the name of the group should be entered.

Online Links

Should be multi-line to allow for more than one link. Links should be clickable. The linked resources could be an online PDF, Google Doc, full website, individual page, forum post, wiki page, YouTube video etc. In the case of some pre-internet languages there will not be an 'official' web page or presence but if there is some page that presents a reasonable amount of information about it that could be used. We may have to consider the necessity of using a book reference instead of a weblink if there is not.

Start Year

The year the creator(s) started creating the language. When the start point is not known, enter the earliest year the language is known to have been in existence.

Physical Mode(s)

What physical media the language is expressed in, such as speech, writing, signing, other rarer media. Options:
1. Speech and writing
2. Speech only
3. Writing only
4. Sign
5. Other

Note: "Speech and writing" should only be used if there is both a 'native' pronunciation and writing system. If the language is only spoken in-world, and a Romanisation or IPA transcription is only for our benefit, that can be assigned to "Speech only".

Scripts

Should be multi-line to allow for more than one entry. Lists the scripts the language uses.

Group

For if the language is in some sort of language group, such as a genealogical one (either a natlang one in the case of altlangs, or a constructed one in a conworld situation) or else a 'suite' in the case of interrelated experimental languages. Can be left blank if the language is in no group.

Type(s)

Classification of the conlang in various types, such as the author's purpose for its creation, its setting, the uses to which it's put etc. As conlangs can be classified in several different types at once this field will allow for multiple choices.

However, the usual top-level categories such as *auxiliary*, *artistic*, *engineered* will not be used; instead the second-order types like fantasy, alternate history, logical, IAL will be the options. There should be guidance on which to pick, and it will be up to database searchers, rather than us designers or contributors/compilers, which second-order types belong to any of the top-level ones. For instance if someone wants to search all artlangs, and considers "personal languages" and jokelangs to be artlangs, they can include those types in the search, along with fantasy, alternate history etc. If they want to search all artlangs but *don't* consider personal languages and jokelangs to be artlangs, they can exclude them from the search. We just leave aside from the database design the question of what counts as what major categories of conlang.

Options:
- "Personal" (created only for the creator's own enjoyment)
- "Jokelang" (created purely for amusement)
- "Story-based" (created to feature in a formal narrative)
- "Conworld" (created as part of an imaginary (constructed) world)
- "Future" (imagined as arising in the future relative to the time of construction)
- "Geofictional" (imagined as being in a fictional part of our world)
- "Alternate History" (imagined as part of an alternative timeline)
- "Lostlang" (imagined as an undiscovered part of our timeline)
- "Exo-/Xeno-lang" (imagined as being used by alien/ET beings)
- "Pseudo-Auxlang" (created to mimic auxlangs but without an intention of auxiliary purpose)
- "Global Auxiliary" (intended for auxiliary use by the whole world)
- "Zonal Auxiliary" (intended for auxiliary use by a limited area of the world)
- "Other Auxiliary" (intended for some other sort of auxiliary use)
- "Ideal" (created to attempt an optimum representation of ideas)
- "Philosophical" (expressing a philosophical viewpoint)
- "Logical" (aiming to express meaning without syntactic ambiguity)
- "Experimental" (testing or demonstrating linguistic possibilities)
- "Conpidgin" (forming a new language by collective evolution)
- "Spiritual/Mystical/Ritual" (inspired mystically or used for ritual purposes)
- "Secret" (stealthlangs, for clandestine communication)
- "Other"

Vocab Source

Classification of the conlang by the origin of its vocabulary. Based on extensive discussion, there should be the following options:

1. 'A priori'/original vocab/made up from scratch/'ex nihilo' - including where the language has been derived from another one the same creator has created in the same constructed family.
2. 'A posteriori'/vocab derived from natlang(s)
3. 'A posteriori'/vocab derived from conlang(s), except where the creator has merely 'derived' one language from another in a family they've created. The conlang copied from should ideally be someone else's creation.
4. A mixture of several of the above
5. Other/Unknown

This issue can be confusing but I don't think we can omit it because it has been a key question in analysis of some kinds of conlang such as auxlangs. It also has some relevance to altlangs. However it would get too confusing to make further distinctions and we'll have to put aside the status of other aspects than vocabulary, such as phonology.

Development Level

How developed the language is, from sketchlang at one end to native speakers at the other. I've added some finer distinctions around the 'developed' area. I'm reluctant to include options like "finished" or "abandoned" because these are not levels of development of the language but current decisions by its creator. Options:

1. "Unknown"
2. "Sketch/Minimal" (either a mere plan or a superficial fictional language)
3. "Some Development" (moving beyond a sketch)
4. "Considerable Development" (halfway to usability)
5. "Well-developed" (the language has been built to a usable level)
6. "Learners" (the language has attracted people to learn it besides its creator)
7. "Active Community" (users are conversing and developing the language)
8. "Fluent Users" (some learners have achieved fluency)
9. "Native Users" (some children have learned the language as a mother tongue)
10. "Other"

Notes

For contributors to add noteworthy information about the conlang that doesn't fit elsewhere. Can contain a "feature summary" or "features of interest". This field should take the form of free text entry.

Of the above, four fields will have a menu of limited options. This will group conlangs of the same kind together, which will help people find them by searching for that kind. The alternative of allowing free text is possible, and would be preferred by some, but would have the negative effect of reducing the accuracy of statistics of each type and returning fewer hits for type searches. There will always be disagreement about how conlangs should be classified, or what terms people use, and the evidence from the spreadsheet shows many creators or other people adding data can be quite individual in how they answer these questions. The rights and feelings of creators should always be considered, but since the purpose of the database is to be accurate and useful we will also have to think in terms of regularising the entries for these fields. However each of them should have an 'Other' option to let contributors avoid having to apply types they don't agree with (or understand). And it's entirely possible that good arguments could lead to us adding new types in some of these fields after the database is started.

What has changed?

The following have been removed:
- The **Lexico-Semantics** field, for classification of conlangs according to the organisation of vocabulary in relation to areas of meaning, e.g. taxonomic vs. naturalistic. This was primarily intended to catch taxonomic languages, but they are a small minority and this is the sort of information that can just be put in **Notes**.
- The original **Purpose Type** field, with top-level options of artistic, auxiliary, engineered etc.

The following have been added:
- The **Conlang Code Registry code**, **Scripts**, **Group** and **Notes** fields.
- More options in the **Purpose Type** (now **Type(s)**) and **Development Level** fields.

The following have been changed:
- The three '**Names**' fields are now multi-line to allow for multiple names.
- Some of the field names have changed, e.g. "**Page**" is now "**Online Links**".
- Some options have been changed for the fields with a limited list, e.g. **Physical Modes** and **Vocab Source**.

There are now 13 fields where there were 10 in the original spreadsheet (plus **Notes** added more recently).

Overall summary of changes:

- This new scheme adds a few more fields, but not many more, and a couple have been removed which makes that easier.
- Suggestions for fields for e.g. previous names or pseudonyms have been accommodated by making the name fields multiple entry rather than single entry. This is neater and will look less complicated on the eventual website data input form.
- More options have been added, and others changed or renamed, in response to suggestions and reminders. At least one multi-option field has been made multiple choice.
- A few have been removed as involving too much complication.
- Field names have been changed to become clearer.

This new scheme is also viewable in its own document: see above in 'Online Locations'.

Conlang source list

We will need to record and work through places to find conlangs and their information. This is the beginning of a list of them, though it will probably have to be added to and revised over time:

Wikis and other documentation websites:
- [CALS](#)
- [FrathWiki](#)
- [Linguifex](#)
- [Langmaker](#) [web.archive]
- [ConWorkShop](#)
- [Conlang Wikia](#)
- [IAL Wikia](#)
- [Anthologica](#)
- [Idéopédia](#) [fr]
- [Geopoeia](#)

Forums and mailing lists:
- [CONLANG mailing list](#) ([Archive](#)/[Mirror](#))
- [AUXLANG mailing list](#)
- [CBB](#)
- [Zompist BB](#) ([up to 2018](#))
- [r/conlangs](#)
- [Constructed Languages – LiveJournal](#)
- [alt.language.artificial](#) [Google Groups]
- [Romance Conlangs list](#) [Yahoo Groups]
- [Germaniconlang](#) [Yahoo Groups]

Lists (Non-personal):
- [List of constructed languages](#) – Wikipedia
- [ConLang Code Registry](#)
- [Conlang Timeline](#) [Land of Invented Languages]
- [Conlanger's Library Timeline Page](#)
- [Auxilingua](#)
- [Auxiliary Conlangs](#) [Auxilingua Documents, web.archive]
- [Linguas](#) [International Auxiliary Languages | Fandom]
- [Kunsttalen](#)
- [CBB Conlang Census](#)
- [ZBB Conlang Index](#)

Lists (Personal):
- [Janko Gorenc's](#)
- [Dean Easton's Page Constructed Languages Library](#) [web.archive]
- [Dirk Elzinga's Constructed Languages page](#) [web.archive]
- [Don Blaheta's Constructed Languages links](#) [web.archive]
- [Garrett's Links to Logical Constructed Languages](#)
- [Joshua Shinavier's Conlang Yellow Pages](#) [web.archive]
- [Mike Ellis' Conlang Links](#)
- [Richard Kennaway's Constructed Languages List](#)
- [Rivendell's Index of Constructed Languages](#)

Books:
- Louis Couturat – Histoire de la Langue Universelle
- Umberto Eco – [The Search for the Perfect Language](#)
- Arika Okrent – [In the Land of Invented Languages](#)
- Stephen Rogers – [Dictionary of Made-Up Languages](#)
- Michael Adams – From Elvish to Klingon: exploring invented languages
- David J. Peterson – [The Art of Language Invention](#)
- Jeffrey Henning – Langmaker: Celebrating Conlangs

Publicity list

Those of us involved in the project need to publicise it around the many conlang communities – the following is a list of such places:

Mailing lists:
- [CONLANG](#)
- [AUXLANG](#)

- [TolkLang](#)
- [Elvish Linguistics List](#)
- [Romance Conlangs](#)
- [Germaniconlang](#)
- maybe some linguistics lists

Forums:
- [CBB](#)
- [Zompist BB](#)
- [L'Atelier Philologique](#) [fr]
- [IdeoLengua](#) [sp]
- [Polskie Forum Językotwórców](#) [pl]
- [Constructed Languages on LiveJournal](#)
- [r/conlangs](#)
- [alt.language.artificial](#)

FB groups:
- [Conlang Conversations & Off-Topic, Also Bhutan](#)
- [Conlang Ideas](#)
- [Conlangs](#)
- [Conlangs and Linguistics, Constructed Languages, Invented Languages](#)
- [Constructed languages](#)
- [Crazy Conlang Ideas](#)
- [Discussing about planned languages: aims and perspectives](#)
- [Linguistics and Conlangs](#)
- [Conlangs (Español)](#) [sp]
- [Germanic and Slavic Conlangs](#)
- [Romance and Celtic Conlangs](#)

AV media:
- [Conlangery Podcast](#)
- some YouTube channels like [Artifexian](#)

Wikis:
- [FrathWiki](#)
- [Linguifex](#)
- [Conlang Wikia](#)
- [IAL Wikia](#)
- [Anthologica](#)
- [Idéopédia](#) [fr]
- [Geopoeia](#)

Lists:
- [Auxilingua](#)
- [Kunsttalen](#)
- [Janko Gorenc's](#)

Other:
- [ConWorkShop](#)