Towards a sustainable handling of interlinear-glossed text in language documentation

JOHANN-MATTIS LIST*, Max Planck Institute for the Science of Human History, Germany
NATHANIEL A. SIMS, The University of California, Santa Barbara, America
ROBERT FORKEL, Max Planck Institute for the Science of Human History, Germany

While the amount of digitally available data on the worlds' languages is steadily increasing, with more and more languages being documented, only a small proportion of the language resources produced are sustainable. Data reuse is often difficult due to idiosyncratic formats and a negligence of standards that could help to increase the comparability of linguistic data. The sustainability problem is nicely reflected in the current practice of handling interlinear-glossed text, one of the crucial resources produced in language documentation. Although large collections of glossed texts have been produced so far, the current practice of data handling makes data reuse difficult. In order to address this problem, we propose a first framework for the computer-assisted, sustainable handling of interlinear-glossed text resources. Building on recent standardization proposals for word lists and structural datasets, combined with state-of-the-art methods for automated sequence comparison in historical linguistics, we show how our workflow can be used to lift a collection of interlinear-glossed Qiang texts (an endangered language spoken in Sichuan, China), and how the lifted data can assist linguists in their research.

CCS Concepts: • **Applied computing** → *Language translation.*

Additional Key Words and Phrases: Sino-Tibetan, interlinear-glossed text, computer-assisted language comparison, standardization, Qiang

ACM Reference Format:
Johann-Mattis List, Nathaniel A. Sims, and Robert Forkel. 2020. Towards a sustainable handling of interlinear-glossed text in language documentation. 1, 1 (March 2020), 15 pages. https://doi.org/+++

1 INTRODUCTION

With many of the world's spoken languages being threatened, efforts on language documentation have been increasing, as reflected in a constantly growing amount of various resources, ranging from short grammatical sketches and wordlists, up to extensive dictionaries, detailed grammars, and corpora in various forms and formats. Depending on the original interests of the researchers, but also on the funding upon which scholars base their research, language documentation follows a range of rather different purposes, as reflected in *typological surveys*, surveys oriented towards *historical language comparison*, *language revitalization efforts*, efforts reflecting *political motives* (such as the dialect surveys

*Corresponding author.

Authors' addresses: Johann-Mattis List, list@shh.mpg.de, Max Planck Institute for the Science of Human History, Kahlaische Str. 10, Jena, Thüringen, 07745, Germany; Nathaniel A. Sims, nsims@ucsb.edu, The University of California, Santa Barbara, 3432 University Drive, Santa Barbara, America; Robert Forkel, forkel@shh.mpg.de, Max Planck Institute for the Science of Human History, Kahlaische Str. 10, Jena, Thüringen, 07745, Germany.

Manuscript submitted to ACM

conducted by Chinese scholars in the 1950s [28]), and efforts reflecting *missionary goals* (such as surveys conducted by religious organizations).

While the amount of digitally available data on the worlds' languages is steadily increasing, with more and more languages being documented, only a very small proportion of the language resources that are produced meet the criterion of *sustainability*. Sustainability – in the context of scientific research – is hereby understood as a resource that complies to the principles of FAIR data as outlined by Wilkinson et al. [35]: resources should be *findable*, *accessible*, *interoperable*, and *reusable*.

The low degree of language resource re-usability is due to the different objectives of people who carry out language documentation. We face a situation where specifically the re-usability of language resources is made extremely difficult. This starts from the fact that some resources are still only produced in print, and even if they are produced digitally, they are rarely *machine-readable*, as they are shared in form of PDF documents, which cannot be converted to computer-friendly resource formats, such as spreadsheet tables or lightweight databases. Even if the data are shared in tabular, basically machine-readable form, they are often not *interoperable*, because they lack *standardization*, and in order to access one specific resource, huge efforts are needed in order to lift the data to a level where they could be easily reused in computer-based or computer-assisted frameworks oriented towards *cross-linguistic comparison*.

One might argue that it is not the primary purpose of language resources, such as, for example, dictionaries, to be parsed by a computer application, but rather by humans who want, for example, to teach an endangered language in school. But it is important to keep in mind that even humans tend to prefer digital dictionaries over resources written in prose and printed only on paper, and the easier a given resource can be searched, the more lasting will be its impact, specifically among younger generations. In addition, the limited sustainability of linguistic resources makes it very difficult, if not even impossible at times, to develop targeted applications in the field of *natural language processing* (NLP), specifically for endangered and poorly documented languages.

Most NLP applications are not only "blind" to language-specific aspects, since – specifically for poorly documented languages – the resources are lacking, but additionally – since large language resources used for the study of big languages (English, Chinese) are often of poor quality – ignore linguistic knowledge to a large degree. In order to side-step the problem of lack of documentation, researchers in NLP now have started to try and impute missing data from cross-linguistic typological databases, given that the data-hungry business of NLP can often not cope with datasets small in size [30]. In fact, prediction (or *retrodiction*) of missing features can indeed be useful, not only in the typological sphere but also for the lexicon, as scholars report in an ongoing experiment of word prediction of Kho-Bwa languages (Tibeto-Burman) [2]. But in order to allow for a successful integration of linguistic resources that could help NLP applications to improve its approaches, specifically also when dealing with smaller and endangered languages, it is important to improve on the general sustainability in language documentation.

While some steps in this direction have been already undertaken in the future, with new standards being proposed for the handling of word lists and structural data in historical linguistics and language typology [6], or initial frameworks having been developed for the handling of rhyme annotation [25], we want to draw the attention to *interlinear-glossed text* as one of the crucial resources produced by language documentation efforts. Although large collections of interlinear-glossed text have been produced so far, and scholars use it across all subfields of linguistics, including opposing camps, the current usage practice largely lacks sustainability, being – despite its formal nature – mostly oriented towards "manual digestion".

In the following, we want to propose a first framework for the computer-assisted, sustainable handling of interlinear-glossed text (IGT). After discussing our general strategy to increase the sustainability of linguistic resources, which

follows closely the recommendations of the Cross-Linguistic Data Formats initiative (https://cldf.clld.org, [6], Section 2), we will present a detailed (but still rudimentary) proposal for the standardization of interlinear-glossed text (Section 2), and illustrate, how this framework can be successfully applied to lift the data of a small corpus of Qiang texts (Section 3), an endangered language, spoken in the northwest part of Sichuan Province in China [11, pp. 1-5]. We conclude by discussing further application possibilities for our framework and point to problems that need to be addressed in the nearer future (Section 5).

2 SUSTAINABILITY OF LINGUISTIC RESOURCES

Given that linguists create linguistic resources with different purposes in mind, the resources – specifically those on endangered and low-resource languages – differ widely. While it is clear that there are generally different type of resources, and that not all linguists plan to create a dictionary of the languages they want to document, the problem does not lie in the broad categories (dictionary, grammar, text corpus, wordlist), but in the way in which the broad categories most scholars would agree upon are created and shared.

As an example, consider the seemingly simple problem of creating *comparative wordlists* for a couple of languages of interest. While the basic format, according to the standard notion of the linguistic sign, would require a triple of *language*, *concept*, and *form*, we find standardization issues in all three of these basic components. Language names, although referring to the same language variety, may vary widely, both for historical reasons (e.g., because language names in the past may have had a derogatory attitude), but also for reasons that are not always made explicit in published studies. Concepts are usually denoted with help of *elicitation glosses*, i.e. the gloss that linguists use in order to elicit a given concept [23], but elicitation glosses that are intended to denote the same concepts vary widely, even if the same language for elicitation has been used [18]. Word forms, finally, are the least standardized of all items one encounters in wordlists, given that scholars usually do not provide phonetic transcriptions, but rather turn to orthographies, where available, or make use of quasi-phonological transcriptions that they consider more convenient for typing, but which are rarely explained with respect to the intended phonetic values.

While the problems may seem severe, initial standardization efforts have been done in the past years, and they have also shown that is possible to successfully enhance existing datasets, by applying a procedure that could be called *retro-standardization*. Instead of changing existing resources manually, semi-automatically, or automatically, retro-standardization adds several annotation layers to existing datasets that allow for an easy conversion of the original data into a format that is machine-readable and cross-linguistically comparable.

These efforts have been most prominently propagated by the Cross-Linguistic Data Formats initiative (CLDF, https://cldf.clld.org, [6]). The basic idea of CLDF is to address comparability problems involving linguistic data by introducing *reference catalogs*, i.e. meta-databases that offer information for those entities which are crucial for cross-linguistic comparison. As the most prominent example, the Glottolog catalog (https://glottolog.org) offers information on language names, geographic locations, and basic genealogical classifications [8]. In order to make sure that it is clear which languages a given resource documents, all that needs to be done is to list the *Glottocodes*, the identifiers provided by Glottolog, for each language that occurs in the resource. Similarly, the Concepticon project (https://concepticon.clld.org, [21], see [20] for details), offers standard identifiers for elicitation glosses and links existing concept lists to those identifiers in order to illustrate the huge variation that can be encountered in concept elicitation. For word forms, the recent Cross-Linguistic Transcription Systems initiative (CLTS, https://clts.clld.org, [19]) provides standard identifiers

for speech sounds which are themselves linked to different transcription systems and thus offer a convenient way to check if a given transcription complies to the standard defined by a given system [1].[1]

CLDF reference catalogs do not stop with providing identifiers to which the original data could be linked. In addition, specific tools are provided that facilitate the process of linking. While identifying languages in Glottolog is already made easy by the web application, the Python API that comes along with it allows scholars proficient in Python programming to use the data provided with Glottolog inside of Python scripts. Concepticon offers command-line tools that allow for an automated mapping of elicitation glosses to the Concepticon identifiers in multiple languages, which can as well be applied from within Python scripts. CLTS offers a range of strategies to normalize transcription data, specifically when provided in the broad version of the IPA that is at the core of the reference catalog. Additionally, scholars can make use of *orthography profiles* [29] that allow for a semi-automated conversion of transcriptions in a given resource into the standards supported by CLTS.[2] All in all, these tools, which are well-documented and also illustrated in several online tutorials, greatly facilitate the process of *retro-standardization* [17]. The advantage of retro-standardization efforts are most prominently illustrated by the Database of Cross-Linguistic Colexifications (https://clics.clld.org, [27]), a large collection of aggregated lexical datasets, which has recently been published in its third version [31], containing data on more than 2400 language varieties, aggregated from 30 different sources.[3]

With respect to interlinear-glossed text, the situation is still different. Although annotation tools exist, as, for example provided by the Summer Institute of Linguistics' FieldWorks program (https://software.sil.org/fieldworks/), their application is difficult due to a lack of cross-platform support (with many tools working only on Windows machines), but also by a large degree of freedom offered by the respective software. Since the majority of IGT is still produced in research articles, and not in form of standardized databases, errors in the glossing procedure are still rather common, as can be seen when checking a random resource provided by ODIN, the largest agglomeration of interlinear-glossed text examples taking from linguistic resources [13].

Our strategy for working towards an increase of sustainability in language documentation, with a specific focus on interlinear-glossed text is two-fold, following the idea of retro-standardization, as it has been proposed by the CLDF initiative. First, we want to increase scholar's awareness regarding available standards and the advantages of using them. Second, we want to make it as easy as possible for scholars to produce their data in the way they know, while encouraging them to open backdoors for quick retro-standardization of their data. The basic idea is to provide initial standards that come close to the formats which scholars already use, but are strict enough to allow for a quick processing by a machine. The advantage of such an approach is that data can be automatically checked for errors which may be easily introduced in typing, while at the same time opening a door for quick retro-standardization with help of computer tools which we will present in detail in the following sections.

3 PROPOSALS FOR STANDARDIZING INTERLINEAR-GLOSSED TEXT

In the following, we will present our proposals for a flexible standardization framework of interlinear-glossed text in detail. After briefly discussing the role that interlinear-glossed text plays in language documentation, we will explain the basic ideas behind the CLDF initiative in more detail, and then present a workflow for the retro-standardization of resources that offer interlinear-glossed text.

[1] The CLTS framework and database is described in detail in the paper by Anderson et al. [1]. Additionally, the supplementary material illustrates in a step-by-step guide how the CLTS code framework can be used for the tasks described in this paper.
[2] See the tutorial by List [16] for details on orthography profiles and how they can be employed. A web-based implementation of the Python package for orthography profiles can be directly accessed and tested at https://digling.org/calc/profile.
[3] By now, CLICS has not only been used in numerous linguistic studies [7, 33], but also in studies dealing with open research questions in psychology [10].

3.1 Inter-linear-glossed text

Inter-linear-glossed text is a commonly used way of presenting the structure by which phrases in languages are built. The basic idea is to gloss each word of a phrase in a certain language by grammatical and lexical glosses in order to elucidate how the respective language expresses a certain circumstance. Technically, IGT demands at least two separators. First, words in the language that is being glossed need to be distinguished, which could be done by a simple white-space character, which is often represented by a tab-stop, in order to support a visual alignment of the original text and the glosses. Second, all meaningful grammatical and lexical units, that is, the *morphemes* inside a word need to be marked, which is usually done with the help of the dash character ("-"). Apart from this, there are different rules to distinguish *lexical* from *grammatical* glosses. The most common way consists in writing grammatical glosses in abbreviated form in capital letters, and providing a legend for the meaning of the abbreviations. Lexical glosses are usually not standardized and simply follow the analysis of the researcher with respect to the utterance under question. Table 1 provides an example of a piece of IGT in German along with the lexical and grammatical glosses and the translation in English.[4]

Die	**Katze**	**sitz-t**	**auf**	**den**	**Matratz-en**
ARTIC.NM.SG.F	cat	sit-3.SG.IND	on	ARTIC.DT.PLR.F	mattress-PLR

The cat sits on the mattresses.

Table 1. Simple example sentence of IGT in German.

Although there have been efforts to standardize IGT with respect to the usage of grammatical glosses, one can encounter a lot of variation with respect to the implementation of the principle. Scholars tend to provide their own abbreviations in the introduction or the appendix of the work, and they also tend to use their own transcription systems (if the language under question has no standardized orthography). Ideally, the information on the grammatical glosses and the transcription systems are exemplified in the studies providing IGT, but the fact that IGT is not following any strict principles – and is barely checked by computational methods for internal consistency – results in a large variation that makes it difficult to make actual use of large IGT collections such as the ones provided, for example, by the ODIN project [13].

While it cannot be denied that there is a certain awareness of the problem of incomparability of IGT from a cross-linguistic perspective, with quite a few journals demanding IGT to follow the popular *Leipzig Glossing Rules* [3], the lack of computer-assisted *testing* whether a given sample of IGT provided in an article or a database conforms to a given standard makes it extremely difficult to compare IGT corpora *across* the studies in which it was originally proposed. Most linguists digest IGT examples piece by piece, without expecting to use them for corpus studies or extended NLP applications. As a result, the majority of IGT corpora produced at the moment is largely incomparable and not amenable for quantitative comparison, at least not beyond the scope of the resource in which they were originally produced. This is extremely unfortunate, given the wealth of information that IGT could offer for cross-linguistic investigations. Although there *are* large resources of digitally available IGT, as it is provided, for example, by the PanGloss project (https://lacito.vjf.cnrs.fr/pangloss/), the Dictionaria project (https://dictionaria.clld.org), or the ODIN corpus [13], there is no way to unify the available resources in a common framework. This is a pity, since IGT offers – at least in theory –

[4]Note that we use English as a glossing language for convenience here. Practically, however, the Concepticon resource is not restricted to any specific glossing language and currently offers active support for many common glossing languages, such as Spanish, Russian, German, French, Portuguese, and Chinese.

many possibilities for interesting analyses that could drastically increase the amount of resources that scholars who work on quantitative applications in NLP, historical linguistics, and linguistic typology have at their disposal. In cases where dictionaries are lacking, one could use larger IGT collections of the same language to construct *wordlists* for cross-linguistic comparison. Where grammatical surveys are lacking, IGT could help to extract *structural features* about a certain language. Finally, if the transcriptions in which IGT is shared were *standardized*, it could give hints not only to *phoneme inventories* but also to the potential usage frequency of the phonemes employed by a given language.

3.2 Workflow for retro-standardization of interlinear-glossed text resources

Our workflow for the retro-standardization of interlinear-glossed text is rather straightforward and seeks to standardize those aspects of a given resource for which reference catalogs as proposed by the CLDF initiative are supported. A minimal example of interlinear-glossed text consists of two entities. First, there is a *text* that is divided into *sentences*, which are themselves divided into *phrases*. Phrases again consist of a sequence of *words* which are themselves divided into *morphemes* (or *morphs*). Second, a sequence of glosses is aligned to the text, with each gloss providing lexical or grammatical semantic information for each morpheme.

While general rules for text glossing have long since been proposed[3], these rules only standardize the outer appearance of interlinear morpheme glossing, while they do not provide any additional recommendations with respect to the way in which, for example, the text should be written, or which elicitation glosses should be used. Since, with the Concepticon project and the CLTS initiative, new reference catalogs are available by now, we think it is time to see to which degree these catalogs can be used to enrich the information that is provided in collections of interlinear glossed text.

Following the general idea of the CLDF initiative of linking resources to the major reference catalogs which have been proposed so far, our workflow towards a retro-standardization of IGT resources thus consists of the following five steps. In a first step, we *standardize* a given IGT resource by making sure that the basic principle of glossing is followed consistently. Starting from a digital IGT resource, we thus check that all *words* in a phrase have at least one *glossed complex* that explains them (1). In a second step, we make sure that each *morpheme* in a word is given a distinct *gloss* (be it grammatical or lexical) (2). In a third step, we try to extract *concept lists* for grammatical and lexical glosses, by creating a *concordance* of each pair of a morpheme and its corresponding gloss in the IGT resource. By automatically distinguishing lexical from grammatical elicitation glosses, this creates two concept lists, one grammatical concept list, and one lexical concept list (3).[5] Having created the concept lists, we try to link the entries in the lexical concept list to the Concepticon resource, and the grammatical concept list to the abbreviations and additional instructions that are usually provided along with a given resource of IGT. In the future, we hope to be able to further link the grammatical glosses to reference catalogs similar to Concepticon, but devoted to abbreviations and elicitation glosses for grammatical concepts in linguistic resources (see, for example, the idea of creating a *Grammaticon* as a counterpart of the Concepticon by Haspelmath [9]). In a fourth step, we try to normalize the transcription system by linking each sound segment that occurs in a given IGT resource to the standard transcription systems (called B(road-coverage)IPA) proposed by the CLTS initiative (4). In a last step, we try to identify *language-internal cognate words* in the IGT resource by clustering all morphemes that show a certain degree of phonetic similarity and are glossed by the same elicitation gloss into the same *word family* (5). The supplementary material offers a detailed example showing how the workflow can be applied to our test datasets.

[5] We are aware that there are borderline cases in which grammatical morphemes cannot be strictly separated from lexical ones. In these cases, we recommend to assign the respective forms both to the lexical and the grammatical concept list.

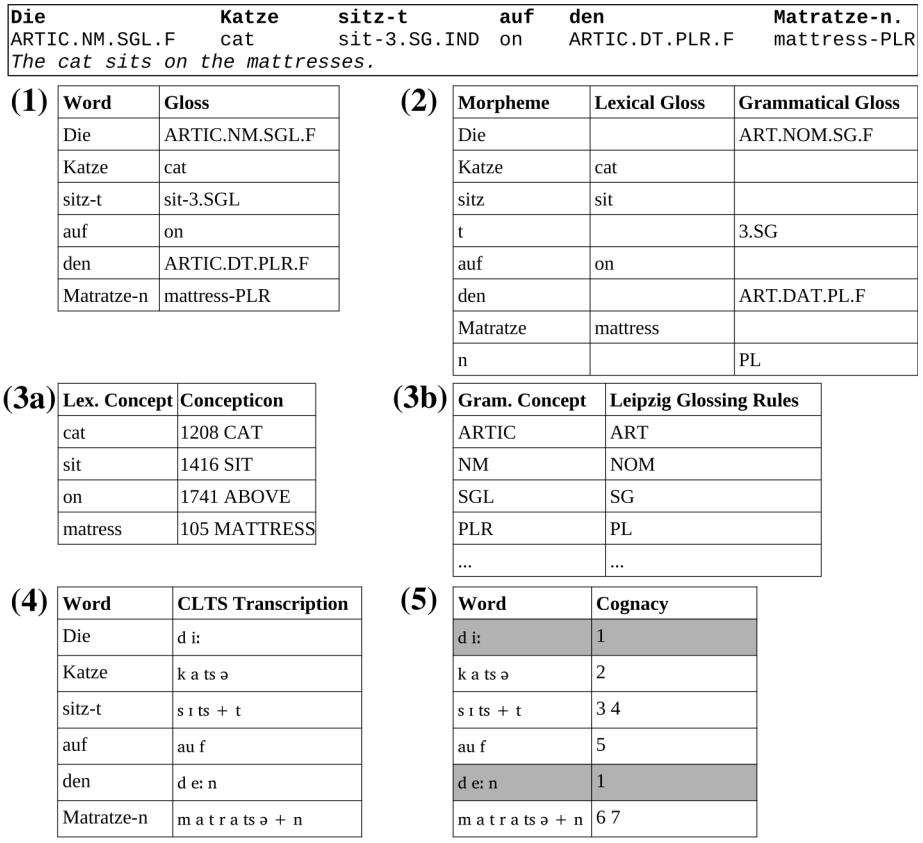

Fig. 1. Five-stage workflow for the normalization of IGT resources. The text example on top of the figure is checked for consistency with respect to words and glosses in (1), and then checked for consistent usage of lexical and grammatical glosses (2). Lexical and grammatical glosses are mapped to Concepticon (3a) and Leipzig Glossing Rules (3b), respectively. All words are transcribed according to the CLTS transcription system (4), and language-internal cognacy is annotated (5).

Once having enriched a given IGT resource in this way, we can present the data in a combined form, in which each instance of the original IGT is accompanied by the additional information that we added during the retro-standardization process. There are two major advantages of this procedure. First, by retro-standardizing data, we increase their cross-linguistic comparability. Increasing comparability also increases the value that a given resource has for the linguistic community, as scholars who are not experts in a specific linguistic area can get quick access to the major information that was accumulated. Second, by applying our procedure for retro-standardization, we check for the internal consistency of the data at the same time. In this way, potential errors in the data can be identified and corrected along with the standardization. While the first advantage may be specifically appealing to typologists, the second aspect is also important for those who collect their data from the field, as it helps to avoid unnecessary errors and inconsistencies, specifically in those cases, where data was collected without assisting software packages. To illustrate how all information acquired with our retro-standardization procedure can be successfully combined, we have created a light-weight web-application in which scholars can query the resource for grammatical and lexical concepts, and word forms. Figure 1 illustrates this workflow in a schematic way.

4 APPLICATION EXAMPLE WITH DATA FROM QIANG (TIBETO-BURMAN)

In the following, we will illustrate how our workflow can be applied to a concrete IGT resource. The supplementary material provides all data and code needed to replicate the experiments we have carried out in this context, but since our work also includes steps of manual refinement, scholars may come to different results when following our example.

4.1 Materials: An interlinear-glossed corpus of Qiang texts

Qiang 羌 (also called Rma) is a Tibeto-Burman language spoken by both ethnic Qiang and ethnic Tibetans in the mountainous area along the upper Min river 岷江 in the Rgnaba-Tibetan-Qiang Autonomous Prefecture of western Sichuan, China. Qiang is not a traditionally written language. It is an endangered language that is in many places being replaced by local varieties of Mandarin [4]. The present Qiang data come from a collection of texts from LaPolla and Huang's 2003 description of the Ronghong variety spoken in northwestern Mao County 茂 [12]. The grammar includes an appendix of six transcribed and annotated texts recorded by three different native speakers. The authors give a free translation into English and Chinese for the texts, but do not provide a line-by-line translation.

In order to make the data amenable for digital treatment, the texts were first digitized and stored in a simple text format which closely renders the format of the glossed text in the original PDF version of the resource, but uses tabstops as standard separators on the word level.

4.2 Methods: A Python package for IGT processing

The code needed to apply the workflow for the retro-standardization of IGT resources is provided in form of a small Python library (*pyigt*), available from the Python Package Index. The code makes use of third-party libraries for a variety of tasks, specifically the LingPy Python library for quantitative tasks in historical linguistics (http://lingpy.org, [22]), which we use in particular for the automated detection of language-internal cognates [24, 26]. The workflow itself is integrated into the CLDF data curation workflow provided by CLDFBench [5], a Python library and command-line tool that facilitates the conversion of different data types into CLDF format. In the following, we will illustrate all steps of our workflow in detail.

4.2.1 Input formats. No specific input formats are needed for the workflow proposed here, since the parsing of the data in their original format and their conversion to the CLDF format for interlinear-glossed text is an integral part of the workflow itself. The format in which the example data are provided in our case is a plain text file, separated into texts, with the transcribed and segmented phrases in one line, and the interlinear glosses in a following line. These two-line blocks themselves are separated by a blank line. From this initial format, we convert the data to CLDF format for interlinear glossed text with help of the CLDFBench package [5]. The workflow itself then uses the data in CLDF format.

4.2.2 Consistency checks on IGT data (1). Once the data is prepared in the format as specified in the preceding section, they can be directly parsed by our library and checked for inconsistencies. During this process, the word forms are also normalized by stripping off punctuation marks and other symbols. This check, which is often only done by eyeballing glossed text resources before publication, turned out to be very useful, since it helped us to identify a couple of inconsistencies in the digital version of the data, which were introduced during the process of digitization.

4.2.3 Creation of lexical and grammatical concordances (2). Once the data has passed the first stage of consistency check, lexical and grammatical concordances can be prepared. In this stage, our workflow checks additionally, if the glosses

match also at the morpheme-level with the words in the resource. In addition, given that grammatical functions often appear in complexes (such as *case*, *number*, and *gender* in many European inflecting languages), this stage introduces a third separator on the level of the glosses, which is used to separate multiple grammatical functions from each other. While the Leipzig Glossing Rules recommend to use a dot for this purpose, the Qiang resource consistently used a colon for this purpose.

The computation of the grammatical and lexical concordances yielded a total of 309 distinct grammatical forms linked to 46 grammatical concepts, and as many as 1201 lexical forms linked to 597 lexical concepts. The most frequently occurring grammatical form was the interjection [ɦa], which we found as many as 360 times in the data, and the most frequently expressed grammatical meaning is represented by numerous directional prefixes (716 examples). The most frequently occurring lexical form was [jə] "say", with 139 occurrences, and the most frequently expressed lexical meaning turned out to be "one" with 206 examples (representing different forms), in many cases used as an indefinite article. All in all, this analysis did not yield any surprises, but it helped us to further eliminate problems in the glosses, as we could identify erroneous glosses that go back to the process of digitization as well as spelling errors in the original resource. An example for a problem in the digitization is the wrong rendering of the word *uncle's* as *unclefls*, which is due to the internal rendering of the apostrophe character in the PDF copy of the grammar. We did not identify many obvious errors (e.g. in spelling) going back to the original source itself, which shows that the resource was thoroughly prepared. An example for a spelling error is the elicitation gloss "daugher" which occurs two times in the original data and obviously refers to "daughter".

4.2.4 Mapping lexical and grammatical concepts to reference catalogs (3). Having extracted lexical and grammatical concept lists, we can *map* the lexical concepts to the Concepticon reference catalog. To ease the mapping procedure, the Concepticon Python API offers an automated mapping routine that checks a given elicitation gloss in a resource against those elicitation glosses that have been used in the 275 resources that have so far been linked to the Concepticon. As a result, the process of concept mapping is greatly enhanced, and it did not take us much time to manually refine the automated mappings.

Having linked the lexical concepts to Concepticon has the advantage of enabling us to check to which degree the concepts in the resource could be used in other applications. Word lists, for example, are important for historical language comparison, but aggregating word lists from different resources is extremely tedious. Once different resources are linked to the Concepticon reference catalog, however, aggregation is simple, since we can automatically check to which degree different resources overlap with respect to the concepts they employ. Thus, of the 591 concepts reflected in the Qiang resource, we find 112 concepts which also occur in the comparative word list collection established by Sagart et al. for their phylogenetic study on Sino-Tibetan languages [32]. A comparison with the concept list of 100 basic vocabulary items proposed by Morris Swadesh [34] shows that the Qiang resource only covers 56 of these concepts. This information is crucial, as it can help scholars who seek to create comparative wordlists from different resources to check quickly if the coverage across different datasets is high enough.

In a similar way, the grammatical concepts offer valuable information, as they can give immediate hints with respect to the grammatical categories which are expressed in a given language. Since no reference catalog for elicitation glosses pointing to grammatical concepts has been established so far, we compared the grammatical concepts in the resource with the list of abbreviations listed in the original resource. In a second step, we added the standard abbreviations suggested by the Leipzig Glossing Rules to the grammatical concept list. While the Qiang resource mostly coincided with the Leipzig Glossing Rules, we find a few interesting cases of divergence. Thus, while the abbreviation PRS is used

by LaPolla and Huang in order to refer to a *prospective aspect suffix*, the abbreviation refers to the *present* tense in the Leipzig Glossing Rules. On the other hand, Lapolla and Huang use *INDEF* to refer to an *indefinite marker*, while the Leipzig Glossing Rules suggest to abbreviate this as *INDF*. While these comparisons may seem pedantic, they greatly exacerbate an automated comparison across resources. Furthermore, the similarity of abbreviations used in different IGT resources but referring to completely different things shows that a careful comparison of linguistic resources can only be done when referring to the original list of abbreviations. In order to guarantee the future comparability of linguistic resources, we need a reference catalog for grammatical elicitation glosses, as well as general efforts to advocate these standards when producing IGT resources.

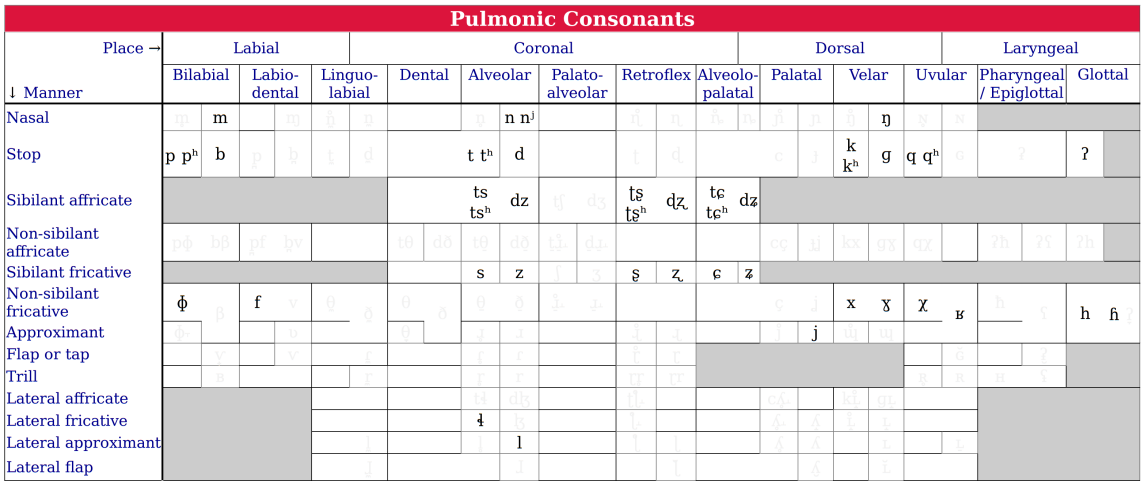

Fig. 2. Consonant chart produced by the EDICTOR tool from the standardized transcriptions.

4.2.5 Standardizing transcriptions (4). As discussed in detail by Anderson et al. [1], the current linguistic practice of phonetic transcription bears not only many pitfalls, but can be barely seen as reflecting a coherent standard. In order to standardize the transcription system employed in a given resource, it is important to identify all distinct sound segments in the data, which can at times be represented by more than just one transcription symbol. While this may sound trivial at first sight, the procedure can turn out to be very tedious, specifically in those cases where a consistent description of the transcription system employed in a given resource is missing.

What has turned out to be extremely helpful in retro-standardizing transcription systems so far is the application of *orthography profiles*, an idea proposed by Moran and Cysouw [29], which consists of a simple table, in which all *graphemes* in a given resource are contrasted with their standardized counterpart. While the original preparation of orthography profiles is tedious, the LingPy software package offers a convenient algorithm for their first creation which also tries to link the transcription symbols to the standard proposed by the CLTS initiative, and which we implemented in our workflow. Once an initial, automated orthography profile has been produced, it can be easily manually corrected.

When adjusting the original transcriptions, it turned out that we did not have to correct many of the transcriptions in the original data. The most notable deviations from the standard transcription system proposed by the CLTS reference catalog was the usage of a normal [h] in order to mark aspiration (which should be represented by a superscript [ʰ]). In addition, we found that the authors often used the letter [a] instead of the letter [ɑ] in order to denote an unrounded

open back vowel, although the former variant is not described in the phonology section of the grammer. We also found instances where orthographical spelling was used instead of the phonetic transcriptions, as in the case of *zz*, which reflects – at least according to the phonological description in the grammar – to a voiced alveolar affricate [dz].

Figure 2 shows a classical IPA chart of all the consonants in the Qiang resource, which was automatically created from the standardized transcriptions with help of the EDICTOR (https://digling.org/edictor/, a web-based tool for the creation of etymological dictionaries [15], which supports the standards proposed by the CLTS reference catalog. As can be seen from this chart, the data does not provide any surprises, but it helps to evaluate a given transcription system and to compare the one we extracted from the glossed texts with the one reported in the grammar.

4.2.6 Identifying language-internal cognates (5). Once created and manually corrected, the orthography profile allows us to convert the original transcriptions into the standardized transcription system and segment the data into sound segments at the same time. This has the great advantage that the data in this form can be easily fed to algorithms for automated sequence comparison as they are provided by LingPy, and as they are needed for the final step of our retro-standardization workflow.

Since IGT resources taken alone never indicate whether two word forms that diverge slightly represent the same lexeme or not, the lexical and grammatical concordances which we created cannot replace a dictionary. What is needed, as a final step, is to make sure that all word forms which stem from the same lexeme, but which differ due to inflection or allomorphic variation, are assigned to the same lexeme entry.

ID	DOCULECT	CONCEPT	CONCEPT TYPE	FORM	TOKENS	OCCURRENCES	WORD FORMS	CROSSID
537	Qiang	market	lexicon	tʂhaq	tʂʰ a/ɑ q	2	tʂhaq tɑ	606
538	Qiang	market	lexicon	tʂhə	tʂʰ ə	1	tʂhə ʐeků tɑ	606
539	Qiang	market	lexicon	tʂhɑq	tʂʰ ɑ q	2	tʂhɑq tɑ	606

Fig. 3. Three slightly diverging word forms denoting "market" in the IGT resource.

In order to identify the lexemes in our data which are reflected by different word forms, we make use of methods for automated sequence comparison in order to produce an initial clustering of similar lexemes into language-internal cognate sets [14]. The result of this analysis is a Qiang wordlist that can be conveniently inspected in the aforementioned EDICTOR tool.

The benefits of this conversion become immediately evident when inspecting the data in detail. As can be seen from the example in Figure 3, we can find three different word forms in the column FORM which all denote the concept "market" in the corpus, which occur together as many as five times. While the two word forms, the first and the third, only differ by their vowel, the second form differs also in the lack of a final consonant. When comparing the differences with our standardized version of the transcription in the field TOKENS, one can see that the difference between [a] and [ɑ] has been accounted for through our orthography profile, in which we already made the decision that [a] is meant to reflect [ɑ]. The segmented form as rendered by the EDICTOR tool still lists this form with a super-script *a*, since we deliberately marked all cases of *a* being meant to represent [ɑ] in our orthography profile.[6] For the form [tʂʰ ə], it is difficult to judge if this is a distinct word or a transcription problem. In any case, what we can clearly see from

[6]This is done by writing the original sound segment and the interpreted sound segment separated by a slash in the replacement column of an orthography profile, thus, underlyingly, the form reads [tʂʰ a/ɑ q] and is rendered as superscript by the EDICTOR.

this example, is, that the procedure of retro-standardizing IGT resources can directly help to improve the resources by pointing to transcription problems.

4.2.7 Exporting the data. As a final step of our workflow, the Python library allows to export the retro-standardized resource to a web-based application that can be used to browse through the IGT examples, searching for lexical and grammatical glosses as well as specific word forms. Given that resources in book form are hard to inspect efficiently, this *concordance browser* offers a very convenient way for typologists and comparative linguists to dive deeper into a given resource. The concordance browser is available from the supplementary material accompanying this study. Figure 4 illustrates its basic usage.

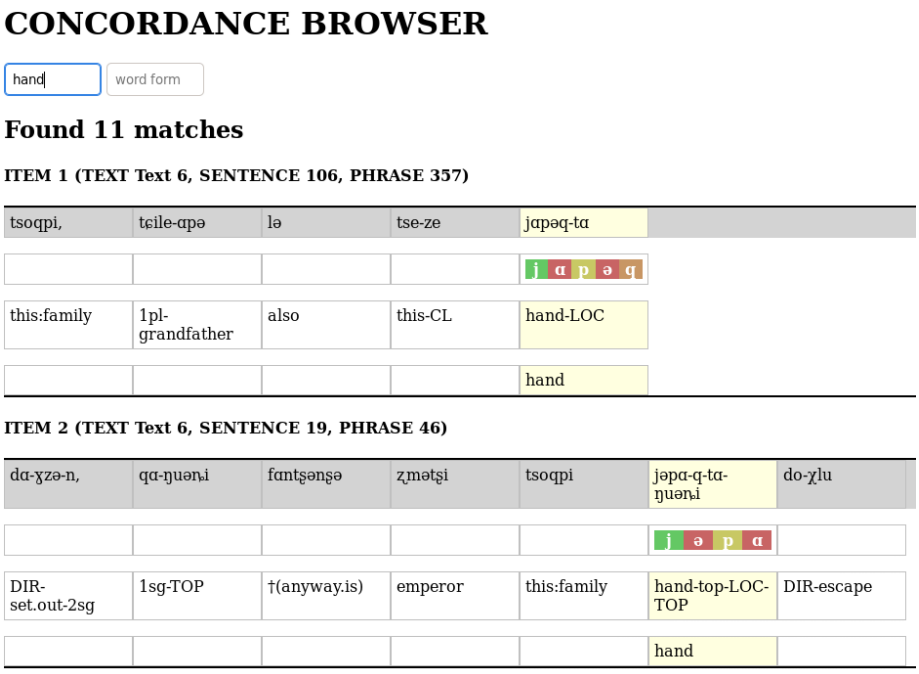

Fig. 4. Searching for occurrences of "hand" in the IGT resources of Qiang with help of the automatically generated *Concordance Browser.*

4.3 Examples

In order to illustrate how the concordance browser constructed from the retro-standardized dataset can be used to shed light on actual linguistic questions, consider the annotation of the hearsay marker [(j)i]. When searching for the grammatical concept "HS", referring to the hearsay marker in Ronghong Qiang, a search with help of the concordance browser yields 24 results, of which the majority of examples has the form [i] (7 occurrences) or [ji] (6 occurrences), as in [oqpi ɦo-pə-i], glossed as `family DIR-become-HS`, which can be translated as "became a family". However, in several of these examples, the form corresponding to the hearsay marker appears as [wei], thus containing a bilabial glide initial which is not present in any of the other examples. While it is difficult to confirm this for all 8 examples it seems there that this form reflects an under-analyzed [-w] morpheme which LaPolla and Huang identify as being part

of the 'non-actor person marking suffixes' elsewhere in their grammar (see e.g., page 120, 143). We therefore think that it is possible that this morpheme is incorrectly being marked as the HS marker, at least in some of the examples, as, for example, in [ʃio-mu-xtɕu-wei], glossed as DIR-NEG-burn-HS, which can be translated as '(they) weren't burned' (Text 1, Phrase 5), or in [de-l-wei], glossed as DIR-give-HS, '(god) gave it to them' (Text 2, Phrase 5).

The analysis of the hearsay marker in the Ronghong variety of Qiang is but one small example of how our retro-standardization can help to shed light on a given IGT resource. If more resources were retro-standardized in the way illustrated here, we think, the great service that interlinear-glossed text provides for typologists and comparative linguistics, can further be increased.

5 OUTLOOK

In this study we have proposed an initial framework for the consistent handling and the retro-standardization of IGT resources in language documentation studies. By illustrating how a concrete resource of a highly endangered Sino-Tibetan language can be successfully retro-standardized and presented in a way that facilitates not only the linguistic but also the computational investigation of the language data, we have tried to show that retro-standardization as well as a sustainable data handling is not *per se* impossible, as scholars often fear, but can even be carried out much more quickly and efficiently than usually assumed. The workflow we propose integrates neatly into previous standardization efforts in the field of computational historical linguistics and computational linguistic typology and requires only a minimal amount of familiarity with the command line in order to be applied successfully.

In the future, we hope to expand our workflow further. First, we want to integrate it more closely with different formats currently used in larger IGT collections, such as PanGloss, ODIN, or the Dictionaria project. Second, we want to discuss with colleagues to which degree it might be possible to establish a reference catalog for grammatical elicitation glosses. Third, given the close integration of this workflow into the CLDF initiative, we want to illustrate and test the usefulness of our workflow by retro-standardizing more datasets and encouraging colleagues to do the same. Last but not least, we want to encourage colleagues working on different languages of the world to test our framework in order to make sure that it is equally well applicable across language families, representing a true, language-independent framework.

SUPPLEMENTARY MATERIAL

The supplementary material contains the source code, the data, and additional instructions on how to use them in order to replicate the analyses discussed here. It is curated on GitHub at https://github.com/cldf-datasets/lapollaqiang and has been archived with Zenodo at https://zenodo.org/record/3626713. The `pyigt` library is also curated on GitHub at https://github.com/cldf/pyigt (Version 0.2) and has been archived with Zenodo at https://zenodo.org/record/3669971.

REFERENCES

[1] Cormac Anderson, Tiago Tresoldi, Thiago Costa Chacon, Anne-Maria Fehn, Mary Walworth, Robert Forkel, and Johann-Mattis List. 2018. A Cross-Linguistic Database of Phonetic Transcription Systems. *Yearbook of the Poznań Linguistic Meeting* 4, 1 (2018), 21–53.

[2] Timotheus A. Bodt and Johann-Mattis List. 2019. Testing the predictive strength of the comparative method: An ongoing experiment on unattested words in Western Kho-Bwa languages. *Papers in Historical Phonology* 4, 1 (2019), 22–44.

[3] Bernard Comrie, Martin Haspelmath, and Balthasar Bickel. 2015. *Leipzig Glossing Rules. Conventions for interlinear morpheme-by-morpheme glosses.* Max Planck Institute for Evolutionary Anthropology, Leizpig. https://www.eva.mpg.de/lingua/pdf/Glossing-Rules.pdf

[4] Jonathan Evans and Jackson T. S. Sun. 2017. Contraction. In *Encyclopedia of Chinese language and linguistics*, Rint Sybesma (Ed.). Vol. 1. Brill, Leiden and Boston, 517–526.

[5] Robert Forkel and Johann-Mattis List. 2020. CLDFBench. Give your Cross-Linguistic data a lift. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*. European Language Resources Association (ELRA), Luxembourg, 1–8.

[6] Robert Forkel, Johann-Mattis List, Simon J. Greenhill, Christoph Rzymski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon A. Kaiping, and Russell D. Gray. 2018. Cross-Linguistic Data Formats, advancing data sharing and re-use in comparative linguistics. *Scientific Data* 5, 180205 (2018), 1–10.

[7] Volker Gast and Maria Koptjevskaja-Tamm. 2018. The areal factor in lexical typology. Some evidence from lexical databases. In *Aspects of linguistic variation*, Daniël Olmen, Tanja Mortelmans, and Frank Brisard (Eds.). de Gruyter, Berlin and New York, 43–81.

[8] Harald Hammarström, Robert Forkel, and Martin Haspelmath. 2019. *Glottolog 4.0*. Max Planck Institute for the Science of Human History, Jena. https://glottolog.org

[9] Martin Haspelmath and Robert Forkel. 2017. Toward a standard list of grammatical comparative concepts: The Grammaticon. Talk held at the database workshop of the ALT Meeting 2017. http://dynamicsoflanguage.edu.au/storage/alt-2017-database-workshop-book-of-abstracts-forkel-haspelmath-haynie-skirgard.pdf

[10] Joshua Conrad Jackson, Joseph Watts, Teague R. Henry, Johann-Mattis List, Peter J. Mucha, Robert Forkel, Simon J. Greenhill, and Kristen Lindquist. 2019. Emotion semantics show both cultural variation and universal structure. Draft article under review. *Science* 366, 6472 (2019), 1517–1522.

[11] Randy J. LaPolla. 1996. *A grammar of Qiang with annotated texts and glossary*. City University of Hong Kong, Hong Kong.

[12] Randy J. LaPolla and Chenglong Huang. 2003. *A grammar of Qiang with annotated texts and glossary*. De Gruyter Mouton, Berlin and New York.

[13] William D. Lewis and Fei Xia. 2010. Developing ODIN: A Multilingual Repository of Annotated Language Data for Hundreds of the World's Languages. *LLC* 25 (2010), 303–319.

[14] Johann-Mattis List. 2014. *Sequence comparison in historical linguistics*. Düsseldorf University Press, Düsseldorf.

[15] Johann-Mattis List. 2017. A web-based interactive tool for creating, inspecting, editing, and publishing etymological datasets. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. System Demonstrations*. Association for Computational Linguistics, Valencia, 9–12.

[16] Johann-Mattis List. 2017. Historical language comparison with LingPy and EDICTOR. https://doi.org/10.5281/zenodo.1042205

[17] Johann-Mattis List. 2017. Historical Language Comparison with LingPy and EDICTOR.

[18] Johann-Mattis List. 2018. Towards a history of concept list compilation in historical linguistics. *History and Philosophy of the Language Sciences* 5, 10 (2018), 1–14. http://hiphilangsci.net/2018/10/31/concept-list-compilation/

[19] Johann-Mattis List, Cormac Anderson, Tiago Tresoldi, Christoph Rzymski, Simon Greenhill, and Robert Forkel. 2019. *Cross-Linguistic Transcription Systems*. Max Planck Institute for the Science of Human History, Jena.

[20] Johann-Mattis List, Michael Cysouw, and Robert Forkel. 2016. Concepticon. A resource for the linking of concept lists. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation*, Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis (Eds.). European Language Resources Association (ELRA), Luxembourg, 2393–2400.

[21] Johann Mattis List, Simon Greenhill, Christoph Rzymski, Nathanael Schweikhard, and Robert Forkel. 2019. *Concepticon. A resource for the linking of concept lists (Version 2.1.0)*. Max Planck Institute for the Science of Human History, Jena. https://doi.org/10.5281/zenodo.3351275

[22] Johann-Mattis List, Simon Greenhill, Tiago Tresoldi, and Robert Forkel. 2019. *LingPy. A Python library for quantitative tasks in historical linguistics*. Max Planck Institute for the Science of Human History, Jena. http://lingpy.org

[23] Johann-Mattis List, Simon J. Greenhill, Cormac Anderson, Thomas Mayer, Tiago Tresoldi, and Robert Forkel. forthcoming. CLICS^2. An improved database of cross-linguistic colexifications: Assembling lexical data with help of cross-linguistic data formats. *Linguistic Typology* 22, 2 (forthcoming). http://clics.clld.org

[24] Johann-Mattis List, Simon J. Greenhill, and Russell D. Gray. 2017. The potential of automatic word comparison for historical linguistics. *PLOS ONE* 12, 1 (2017), 1–18.

[25] Johann-Mattis List, Nathan W. Hill, and Christopher J. Foster. 2019. Towards a standardized annotation of rhyme judgments in Chinese historical phonology (and beyond). *Journal of Language Relationship* 17, 1 (2019), 26–43.

[26] Johann-Mattis List, Philippe Lopez, and Eric Bapteste. 2016. Using sequence similarity networks to identify partial cognates in multilingual wordlists. In *Proceedings of the Association of Computational Linguistics 2016 (Volume 2: Short Papers)*. Association of Computational Linguistics, Association of Computational Linguistics, Stroudsberg, 599–605.

[27] Johann-Mattis List, Christoph Rzymski, Simon Greenhill, Tiago Tresoldi, and Robert Forkel. 2019. *CLICS: Database of Cross-Linguistic Colexifications*. Max Planck Institute for the Science of Human History, Jena. http://clics.clld.org/

[28] Anatole Lyovin. 1969. Review of Hànyǔ fāngyīn zìhuì by Běijīng Dàxué. *Language* 45, 3 (1969), 687–697. http://www.jstor.org/stable/411456

[29] Steven Moran and Michael Cysouw. 2018. *The Unicode Cookbook for Linguists: Managing writing systems using orthography profiles*. Language Science Press, Berlin. http://langsci-press.org/catalog/book/176

[30] Yugo Murawaki. 2019. Bayesian learning of latent representations of language structures. *Journal of Computational Linguistics* 45, 2 (2019), 199–228. https://doi.org/10.1162/COLIa00346

[31] Christoph Rzymski, Tiago Tresoldi, Simon Greenhill, Mei-Shin Wu, Nathanael E. Schweikhard, Maria Koptjevskaja-Tamm, Volker Gast, Timotheus A. Bodt, Abbie Hantgan, Gereon A. Kaiping, Sophie Chang, Yunfan Lai, Natalia Morozova, Heini Arjava, Nataliia Hübler, Ezequiel Koile, Steve Pepper, Mariann Proos, Briana Van Epps, Ingrid Blanco, Carolin Hundt, Sergei Monakhov, Kristina Pianykh, Sallona Ramesh, Russell D. Gray, Robert Forkel,

and Johann-Mattis List. 2020. The Database of Cross-Linguistic Colexifications, reproducible analysis of cross- linguistic polysemies. *Scientific Data* 7, 13 (2020), 1–12.

[32] Laurent Sagart, Guillaume Jacques, Yunfan Lai, Robin Ryder, Valentin Thouzeau, Simon J. Greenhill, and Johann-Mattis List. 2019. Dated language phylogenies shed light on the ancestry of Sino-Tibetan. *Proceedings of the National Academy of Science of the United States of America* 116 (2019), 10317–10322. Issue 21.

[33] Antoinette Schapper. 2019. The Ethno-Linguistic Relationship between Smelling and Kissing: A Southeast Asian Case case-study. *Oceanic Linguistics* 58, 1 (2019), 92–109.

[34] Morris Swadesh. 1955. Towards greater accuracy in lexicostatistic dating. *International Journal of American Linguistics* 21, 2 (1955), 121–137. arXiv:1263939

[35] Mark D. Wilkinson, Michel Dumontier, Ilsbrand J. Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz B. da Silva Santos, Philip E. Bourne, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3 (2016), 1–8.